



Models for efficient integration of solar energy

Bacher, Peder

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):
Bacher, P. (2012). *Models for efficient integration of solar energy*. Technical University of Denmark. IMM-PhD-2012 No. 272

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Models for efficient integration of solar energy

Peder Bacher

DTU



Kongens Lyngby 2012
IMM-PhD-2012-272

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-PhD-2012-272

Summary (English)

Efficient operation of energy systems with substantial amount of renewable energy production is becoming increasingly important. Most renewables are dependent on the weather conditions and are therefore by nature volatile and uncontrollable, opposed to traditional energy production based on combustion. The "smart grid" is a broad term for the technology for addressing the challenge of operating an energy system, especially the electrical grid, with a large share of renewables. The "smart" part is formed by technologies, which model the properties of the systems and efficiently adapt the load to the volatile energy production using the available flexibility in the system.

In the present thesis methods related to operation of solar energy systems and for optimal energy use in buildings are presented. Two approaches to forecasting of solar power based on numerical weather predictions (NWP) are presented. They are applied to forecast the power output from PV and solar thermal collector systems. The first approach is based on a developed statistical clear-sky model, which is used for estimating the clear-sky output solely based on observations of the output. This enables local effects such as shading from trees to be taken into account. The second approach to solar power forecasting is based on conditional parametric modelling. It is well suited for forecasting of solar thermal power, since it can be made non-linear in the inputs. The approach is also extended to a probabilistic solar power forecasting model.

The statistical clear-sky model is furthermore used as basis for a method for correction of global radiation observations. This method can be used for correction of typical errors, for example from shading trees or buildings.

Two methods for flexible heat use in buildings are presented in the last part of the thesis. First a method for forecasting of the heat load in single-family houses based on weather forecasts is presented. A model is identified for forecasting the heat load for sixteen single-family houses. The model adapts to the individual houses and needs no specific information about the buildings. Finally, a procedure for identification of a suitable model for the heat dynamics of a building is presented. The applied models are grey-box model based on stochastic differential equations and model identification is carried out with likelihood ratio tests. The models can be used for providing detailed information of the thermal characteristics of buildings and as basis for optimal control for flexible heating of buildings.

Summary (Danish)

Energiproduktion med vindmøller og solceller kan af naturlige årsager ikke styres som energiproduktion baseret på forbrænding. Derfor kræves en udvikling af nye teknologier til styring af energisystemer med en betydelig andel vind- og solenergi. Elnettet skal udbygges med it-teknologi og blive til et "smart grid". Dette betyder at styringen skal udføres automatisk baseret på computermodeller, der "lærer" –at tilpasse sig systemets egenskaber udfra data opsamlet af sensorer. I afhandlingen præsenteres en række af sådanne modeller og metoder relateret til solenergi og opvarmning i bygninger.

Først præsenteres to metoder til forudsigelse af solenergiproduktionen udfra meteorologiske vejrprognoser. De kan bruges til forudsigelse af energiproduktionen fra både solceller og solfangere, og kræver ikke nogen specifik information om anlæggene udover målinger af energiproduktionen. Den ene metode er baseret på en udviklet statistisk model, der kan estimere solenergiproduktionen ved klar himmel, udelukkende ud fra målinger af produktionen. Dette gør modellen i stand til at inkludere lokale effekter, som for eksempel solenergi-anlæggets vinkel og skygger fra objekter i de lokale omgivelser, f.eks. fra træer. Den anden metode er baseret på en betinget parametrisk model, hvilket giver en effektiv modellering med ikke-lineære funktioner. Denne metode er desuden velegnet til solvarme forudsigelser. Endeligt beskrives en metode til probabilistisk forudsigelse af solenergiproduktion.

En metode til korrektion af fejl i målinger af global stråling præsenteres. Den statistiske model for klar himmel benyttes her på målinger af global stråling og systematiske fejl korrigeres, som for eksempel skyldes skygger fra træer eller bygninger.

En række modeller af solfangeres varmedynamik præsenteres. Modellerne er baseret på stokastiske differentialligninger og kan benyttes til estimation af vigtige fysiske parametre ud fra målinger fra en enkelt dag. Dette gør dem oplagte til brug i tests af ydeevne.

To metoder der kan bruges i styringen af fleksibelt varmekonsum i bygninger præsenteres i sidste del af afhandlingen. Først identificeres en model til forudsigelse af varmekonsum baseret på meteorologiske forudsigelser, og den afprøves på seksen typiske parcelhuse op til 42 timer frem i tiden. Modellen tilpasser sig automatisk de enkelte huses varmekonsum og kræver ingen specifik information om de enkelte huse.

Til sidst præsenteres en metode til identificering af en velegnet model af bygningers varmedynamik. Modellerne er baseret på stokastiske differentialligninger og kan bruges til at estimere vigtige fysiske parametre, for eksempel bygningens UA-værdi og varmekapacitet. De kan for eksempel benyttes til at bestemme potentialet for energirenovering og som basis for styring af fleksibelt varmekonsum.

Preface

This thesis was prepared at the department of Informatics and Mathematical Modelling (Informatics), the Technical University of Denmark (DTU) in partial fulfilment of the requirements for acquiring the Ph.D. degree in engineering. The Ph.D. project was carried as part of the project "Solar/electric heating systems in the future energy system" which is funded by The Danish Council for Strategic Research.

The thesis deals with modelling for operation of energy systems depending on solar radiation and for flexible heating of buildings. Methods for forecasting of solar power and for correction of observations are presented, together with models for forecasting of heat load in buildings and models for the heat dynamics of buildings.

The thesis consists of a summary report and seven research papers. Two of the papers are published in, and two are submitted to, international peer reviewed scientific journals. The last three appear in conference proceedings.

Lyngby, 2012

Peder Bacher

Acknowledgements

My thanks first of all goes to my supervisors Henrik Madsen and Henrik Aalborg Nielsen for being visionary and inventive, and to inspire through a sincere devotion to research and new ideas. Thanks to Bengt Perers for that heartfelt sense of dedication to solar energy research and many eye-opening discussions.

Thanks to all my colleagues at IMM, especially you lovely fellow students for creating an inspiring and really "we get further if we help each other and have fun while doing it"-atmosphere. Very special thanks my office fellows Pierre-Julien and Philip for being so good and meaningful friends. And Janne, thanks for keeping track of things, without you we would not get far before chaos rises to above our ears.

Thanks to all the project participants from DTU Byg and DMI for the fruitful sharing of insights and competencies, especially Simon Furbo for leading the project and bringing it all together. Special thanks to Rasmus Halvgaard for using the forecasts as basis for his work on optimized control of solar and electric heating systems.

All my remaining thanks goes to Christina and our two boys Carl and Vilfred for all the love and support you have given me.

List of publications

Papers included in the thesis:

- A** Peder Bacher, Henrik Madsen, Bengt Perers, and Henrik Aalborg Nielsen. A non-parametric method for correction of global radiation observations. Submitted to: *Solar Energy*, (April 2012).
- B** Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. Online Short-term Solar Power Forecasting. *Solar Energy*, 2009, 83(10), pp. 1772-1783.
- C** Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. Online Short-term Solar Power Forecasting. *1st International Workshop on the Integration of Solar Power into Power Systems. Aarhus, 2011.*
- D** Peder Bacher, Henrik Madsen, and Bengt Perers. Short-term solar collector power forecasting. *ISES Solar World Congress 2011.*
- E** Peder Bacher, Henrik Madsen, and Bengt Perers. Models of the heat dynamics of solar collectors for performance testing. *ISES Solar World Congress 2011.*
- F** Peder Bacher, Henrik Madsen, Henrik Aalborg Nielsen, and Bengt Perers. Short-term heat load forecasting for single family houses. Submitted to: *Buildings and Energy*, (April 2012).
- G** Peder Bacher and Henrik Madsen. Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings*, 2011, 43(7), pp. 1511-1522.

Contents

Summary (English)	i
Summary (Danish)	iii
Preface	v
Acknowledgements	vii
List of publications	ix
1 Introduction	1
1.1 Solar energy	3
1.2 Buildings and energy	7
2 Methods for solar energy applications	11
2.1 Statistical clear-sky model	11
2.1.1 Correction of global radiation observations	13
2.2 Solar power forecasting	15
2.2.1 Two-stage method based on the statistical clear-sky model	15
2.2.2 Solar power forecasting with a conditional parametric model	16
2.2.3 Probabilistic solar power forecasting	16
2.3 Modelling the heat dynamics of solar thermal collectors	21
2.4 Discussion	23
3 Methods for building energy applications	25
3.1 Heat load forecasting	25
3.2 Models for the heat dynamics of buildings	28
3.3 Discussion	31

4	Discussion and conclusion	33
4.1	General discussion	33
4.2	Conclusion	35
	Bibliography	37
A	A non-parametric method for correction of global radiation observations	43
A.1	Introduction	45
A.2	Data: Observations and numerical weather predictions of global radiation	47
A.2.1	Observations	47
A.2.2	Numerical weather predictions	49
A.2.3	Systematic errors in Sønderborg observations	49
A.3	Statistical clear-sky model	51
A.4	Correction of observations	56
A.4.1	On-line operation	58
A.5	Discussion	59
A.6	Conclusion	60
A.7	Two-dimensional local statistical clear-sky model	61
B	Online Short-term Solar Power Forecasting	65
B.1	Introduction	66
B.2	Data	69
B.3	Clear sky model	71
B.4	Prediction models	77
B.4.1	Transformation of NWP's into predictions of normalized solar power	77
B.4.2	AR model identification	78
B.4.3	LM_{nwp} model identification	80
B.4.4	ARX model identification	80
B.4.5	Adaptive coefficient estimates	82
B.5	Uncertainty modelling	82
B.6	Evaluation	83
B.6.1	Error measures	84
B.6.2	Reference model	84
B.6.3	Results	85
B.7	Conclusions	88
B.8	Weighted quantile regression	88
B.9	Recursive least squares	89

C	Online Short-term Solar Power Forecasting	93
C.1	Introduction	94
C.2	Data	95
C.2.1	Pre-processing	95
C.3	Clear sky model	96
C.3.1	Statistically estimated clear sky solar power	97
C.4	Forecasting models	98
C.4.1	Reference model	98
C.4.2	Autoregressive models	99
C.4.3	Conditional parametric models	99
C.4.4	Autoregressive model with exogenous input	100
C.5	Evaluation	102
C.5.1	Error measures	102
C.5.2	Completeness	103
C.6	Results	103
C.7	Discussion and applications	104
C.8	Conclusion	105
D	Short-term solar collector power forecasting	107
D.1	Introduction	108
D.2	Data	110
D.2.1	Solar power	110
D.2.2	Numerical weather predictions	110
D.2.3	Pre-processing	111
D.3	Clear sky model	112
D.3.1	Statistically estimated clear sky solar power	113
D.4	Forecasting models	114
D.4.1	Reference model	114
D.4.2	Autoregressive models	115
D.4.3	Conditional parametric models with NWP's as input	115
D.4.4	Autoregressive model with exogenous input	117
D.4.5	Combined model	119
D.5	Evaluation	119
D.5.1	Error measures	119
D.5.2	Completeness	120
D.6	Results	120
D.7	Discussion and applications	122
D.8	Conclusion	123
E	Models of the heat dynamics of solar collectors for performance testing	125
E.1	Introduction	126
E.2	Grey-box models of a dynamic system	129
E.2.1	Maximum likelihood estimation of parameters	130

E.3	Experimental setup and data	131
E.4	Multiple linear regression models	132
E.5	Applied grey-box models	132
E.5.1	Models with multiple compartments in the flow direction	135
E.5.2	Models divided into a collector and a fluid part	136
E.6	Results	137
E.6.1	MLR models	137
E.6.2	ToComp1 fitted to 10 minutes values	138
E.6.3	Grey-box models fitted to 30 seconds values	139
E.7	Discussion and applications	143
E.7.1	Applications	143
E.8	Conclusion	144
F	Short-term heat load forecasting for single family houses	147
F.1	Introduction	148
F.2	Data	150
F.2.1	Heat load measurements	150
F.2.2	Local climate observations	151
F.2.3	Numerical weather predictions	152
F.2.4	Combining local observations with NWP's	153
F.3	Models	153
F.3.1	Time adaptive models	154
F.3.2	Diurnal curve	154
F.3.3	Low-pass filtering for modelling of building dynamics	154
F.3.4	Parameter optimization	155
F.4	Model identification	156
F.4.1	First step in model selection	156
F.4.2	Second step in model selection	159
F.4.3	Step three: Inclusion of wind speed in the model	161
F.4.4	Step four: Enhancement of the solar model part	162
F.5	Noise model	164
F.6	Results	165
F.6.1	Model parameters	166
F.6.2	Forecasting performance	167
F.7	Discussion	169
F.8	Conclusion	170
G	Identifying suitable models for the heat dynamics of buildings	175
G.1	Introduction	177
G.2	Grey-box models of a dynamic system	178
G.2.1	Maximum likelihood estimation of parameters	180
G.3	A statistical test for model selection	180
G.3.1	Likelihood ratio tests	180
G.3.2	Forward selection	181

G.4	Model selection procedure	181
G.4.1	Model selection	182
G.4.2	Model evaluation	183
G.5	Case study: model identification for a building	184
G.5.1	Description of the building and measurement equipment	184
G.5.2	Data	185
G.5.3	Applied models	186
G.5.4	Model identification	191
G.5.5	Model evaluation	192
G.6	Applications	198
G.7	Conclusion	199
G.8	RC-networks of applied models	200
G.9	Estimates of system and observation noise parameters	203

CHAPTER 1

Introduction

The foreseen negative impact on human life conditions due to climate changes caused by the increasing level of greenhouse gases in the atmosphere, is driving a transformation of the energy system: away from the current system based on fossil fuels and over to a future system based on sustainable and climate neutral energy production. Many plans for this transformation have been presented during the last decades and the visions are manifold. They range from focusing on yet not operational technologies, such as nuclear fusion energy, and over to focusing on existing technologies, such as nuclear fission energy, and wind and solar energy. Reality is however, that no single sustainable energy technology, which can cheaply replace coal and oil, and cover the entire demand for energy, has been invented. Currently, it seems like a very realistic scenario is a system composed of wind, solar, bio, and hydro power. See for example the scenarios suggested by Fthenakis et al. (2009), Jacobson and Delucchi (2011), and Mathiesen et al. (2011). The transition should be carried out with the socio-economically most optimal solutions and is highly dependent on the current energy infrastructure and locally available resources. Due to the spacial requirements and intermittent nature of wind and solar energy technologies, two very important aspects have to be considered: energy storage and energy mobility. In one end of the renewable energy scenarios, all energy is produced with large scale plants at locations where the resource is plentiful: solar energy in the desert and wind energy off-shore at sea. This requires strong interconnections over very long distances for transportation of energy to where it is

needed. In the other end, scenarios are suggested where the energy is produced on distributed small scale systems as close as possible to where it is needed. For the Danish energy system several plans for a system based on 100% renewables in 2050 have been put forward. The currently most established plan is put forward by the Danish Commission on Climate Change Policy (2010), who suggest a profound change of the structure of energy system. Away from the current centralized system, which is based on non-volatile power generation technologies and where the power generation is controlled to match the demand, and over to a distributed system, which can handle a much more volatile power generation and where the demand is shifted to match the power generation. The change involves engineering challenges in many fields, where especially the "smart grid" is pointed out as an essential component which must be developed.

The "smart grid" is a term covering a range of technologies for enabling an efficient and flexible operation of a renewable based energy system, in particular with the use of information and communication technology for providing the needs for communication between the units in the system. This forms the basis for the "smart" part, namely the operation of the energy system with computer models, which can "learn" the characteristics of the system and enable an optimized use of the vast amounts of data from sensors in the system. In the present thesis such methods based on such data-driven statistical models are presented. They are centered around the use of real-time energy and climate data for modelling, with focus on solar energy applications and optimal energy use in buildings. This type of methods will be vital for operation of energy systems with a substantial amount of solar energy production, and - especially for the methods related to buildings - also for efficient integration of other renewables, in particular wind energy. Methods for the following applications are presented in the thesis:

- Solar power forecasting, both for PV systems and solar thermal systems.
- Processing and correction of solar radiation measurements.
- Dynamical models for solar thermal collectors.
- Forecasting of heat load for buildings.
- Identification of suitable models for the heat dynamics of buildings.

The methods are based on statistical modelling techniques, which, where it is appropriately, are combined with prior physical knowledge. The models deal with both non-linear and dynamic relations, and thorough statistical analysis for evaluation of the results are carried out. An objective is to provide a solid foundation for operational applications, as well as expanding current state of the art in the addressed fields. Data from both experimental setups and systems

in operation is used, for example data consisting of measurements from test buildings, inhabited buildings, weather stations, and meteorological forecasts.

The PhD-study is carried out under the project "Solar/electric heating systems in the future energy system", in which a heating system scaled for a single-family house is designed, see (Perers et al., 2011) for a detailed description. The core of the system is a hot-water tank, which is connected to both an electrical heater and a solar thermal collector. The system provides both space heating and hot water heating. By heating in advance with electricity, and storing thermal energy in the tank, the system can be used for load-shifting, and thus become valuable also for the integration of wind power (Meibom et al., 2007). An objective of the PhD-study is to provide the forecasting models needed for optimal operation of the system in a Danish context, where the electricity is bought at a variable price on the Nord Pool Spot day-ahead market. The optimal operation is carried out by shifting the load to hours with cheap energy in order to minimize the costs of running the system. The presented methods for forecasting of the hourly solar thermal production and forecasting of heat load, are used to provide the necessary basis for the economic model predictive control (MPC) scheme presented by Halvgaard et al. (2012) for optimal operation of solar-electric combisystems. Furthermore price forecasts are used, which are provided using the models described by Jónsson et al. (2012).

1.1 Solar energy

Energy from solar radiation drives the processes that makes life on earth possible. It has the potential to cover all human energy needs: the energy from solar radiation striking the surface of the earth in 90 minutes (appr. $6.4 \cdot 10^{20}$ J) is well over the current global energy consumption per year (appr. $5.1 \cdot 10^{20}$ J in 2009 (IEA, International Energy Agency, 2011b)). Several plans exist for a renewable based energy system relying heavily on solar energy. For example: Zweibel et al. (2008) plan 35% of the total energy to come from solar, Heide et al. (2010) finds that an optimal balance is 55% wind and 45% solar energy for Europe, and IEA, International Energy Agency (2011a) discuss the IEA High-Renewable Scenario where 25% of global electricity is covered by solar power. A vast palette of technologies exist for using solar radiation as an direct source of energy: from agriculture over architecture and over to conversion of solar radiation directly to electrical power. Many other renewable energy sources are indirectly driven by solar radiation, for example wind and hydro power. The most used technologies, which convert solar radiation directly into heat or electricity, can be divided into following categories:

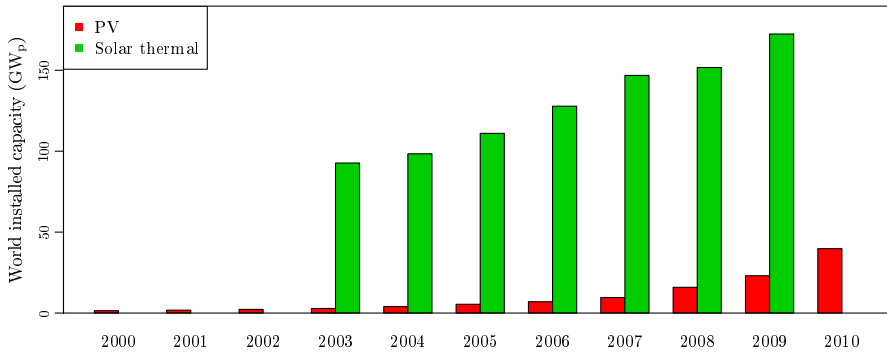


Figure 1.1: Total worldwide installed solar PV and thermal capacity

- Heating, cooling and ventilation in buildings, for example there is a long tradition for buildings to have windows for providing indoor light and heating.
- Water heaters. Different types of solar thermal collectors exist, the most widespread type is flat-plate collectors.
- Photovoltaic (PV) panels, which are either distributed on roof-tops or installed in large power plants, currently the largest is Golmud Solar Park (200 MW_p), which is to be built in China.
- Concentrated solar power. Installed in large plants, where sun beams are concentrated with mirrors to heat fluid for driving steam turbines. This technology has a huge potential in sunny regions, currently the largest is Brightsource's Ivanpah plant in the US, which is planned to reach 392 MW_p when completed. The use of a heated fluid in the process has the advantage that thermal storage can be used to shift the electricity production.

The plot in Figure 1.1 shows the total worldwide installed capacity over the last years of PV (BP, 2011) and solar thermal (appr. 85% from flat-plate and evacuated tube collectors and 15% from unglazed collectors) (IEA - The Solar Heating and Cooling Programme, 2010). Clearly the installed capacity is rapidly increasing and especially the installed PV capacity has been rapidly increasing the last few years. Currently Germany has the highest level of solar power penetration at around 3% of the electricity production, which already has a considerable impact on the power price on sunny days (Nicolosi and Nabe, 2011). The currently installed capacity in Denmark is around 22.5 MW_p, which corresponds to 0.06 % of the electricity production. This is quite little, but in 2011 nearly 8.5 kW_p was installed and the trend seems to be continuing upwards, as seen from the

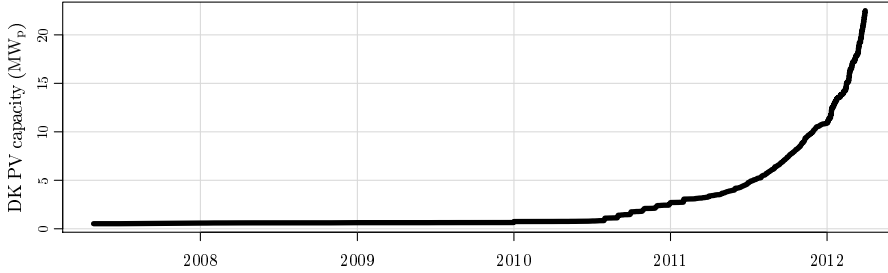


Figure 1.2: Total installed solar PV peak power capacity in Denmark (source: Energinet.dk)

plot in Figure 1.2 of the total installed PV capacity in Denmark. During the large scale smart grid research project Ecogrid, the Danish island of Bornholm (population around 42000) will be equipped with 5MW_p solar power, which will be around 9% peak power penetration (Yang et al., 2011). Hence there is a growing need for methods for an efficient integration of PV solar power.

As seen in Figure 1.1 the globally installed solar thermal capacity is much higher than the installed capacity of PV. The by far largest part of solar thermal is used for water heating and works independently of the surrounding energy system. However there is an increasing trend in application of solar heating for other purposes, such as space heating and industrial heating. As an example is the solar heating system from Innogie¹, where the entire roof is turned into a solar thermal collector. It is coupled to a hot water tank and a geothermal heat pump. In the summertime an HP/ORC module is used for conversion of the solar thermal energy to electricity. Apart from efficient use of solar and electrical energy for heating, the system can also provide flexible electrical load and production.

In order to operate systems with solar power optimally, it is essential to have forecasts of solar power available (Lew et al., 2010). They are required for periods ranging from days ahead down to hours and ten-of minutes ahead (Sayeef et al., 2011). Quite a lot of research in solar forecasting has been carried out the last years and two main approaches, depending on the forecast horizon, have had most attention: for short horizons within a couple of hours, the forecasts are based on satellite images and total sky imagery, and for longer horizons up to a several days, the forecasts are based on numerical weather predictions (NWP).

The following is a small overview of the recently presented methods for forecasting based on NWP. Ji and Chee (2011) do a detrending of solar power

¹www.innogie.dk

by fitting a high order polynomial model based on a monthly average of the diurnal curve to account for non-stationarity of the observed process. Then the detrended series is forecasted with an ARMA and TDNN hybrid model. Lorenz et al. (2011) use several steps in a post-processing procedure of NWP to derive optimized site-specific irradiance forecasts. First a spatial averaging of the NWP is carried out and thereafter a bias-removal procedure, which is based on a physical clear-sky model, is applied. Finally, a physical model is used to convert the irradiance forecasts to PV power forecasts. An up-scaling method for regional solar power forecasting is also presented. Schmelter et al. (2011) presents a continuous weather classification for combining NWP, which improves forecasting performance especially for difficult weather conditions such as fog and snow. Pelland et al. (2011) do post processing of NWP for forecasting global radiation and PV output. First a spacial averaging of the NWP is applied and a Kalman filter is used for bias removal. Fonseca Júnior et al. (2011) compare neural networks and support vector regression for forecasting 24 hours ahead. It is found that RMSE performance is quite equivalent, while MAE performance is better for support vector regression.

In the present thesis two approaches to solar power forecasting based on NWP are presented. The first approach is based on a developed statistical clear-sky model, which is used for normalization of the solar power to a stationary process, such that classical linear time series models, here ARX models, can be efficiently applied. The second approach is to forecast the solar power directly with a conditional parametric model. This provides a simple and efficient approach to solar power forecasting. Furthermore an outline of an approach to probabilistic solar power forecasting is given.

In addition to the forecasting methods two other solar energy related studies are presented in the thesis. First a method for correction of global radiation observations based on the statistical clear-sky model is presented. The method can especially be useful for correction of observations from partly unsupervised solar radiation sensors, which will be exposed to many types of errors, for example shading from surrounding objects. Finally, a study, in which grey-box models for the heat dynamics of solar thermal collectors are applied, is presented. The models can be used for obtaining detailed knowledge of the heat dynamics of a collector, including the physical parameters related to the performance of the collector.

1.2 Buildings and energy

An essential key to a successful integration of the volatile energy production from renewables is the development of cheap and operational energy storage. Today many different energy storage technologies exist. They are different with regards to several aspects such as: efficiency, temporal characteristics, and geographic requirements. For example pumped hydro, which can provide seasonal storage with a round-trip efficiency up to 85%, has a limited potential due to the geographic requirements. Another very promising technology is renewable power methane (RPM) storage, which is based on conversion of electricity to methane gas. The gas is stored and can be used later in the existing natural gas energy infrastructure. The drawback of RPM is that the round-trip efficiency is only in the range of 30% to 37% (Breyer et al., 2011), which makes it expensive and feasible mostly for seasonal storage. Hence cheap and efficient short-term energy storage is really valuable, in Denmark especially for matching the fluctuations in wind power (Meibom et al., 2007), but also for shifting the load away from peak hours within the day. Due to the generally increasing electricity load, even very short time flexibility can be very valuable in order to avoid overload of the transmission system and thereby minimize the needed investments in the transmission system (Danish Commission on Climate Change Policy, 2010). The use of electric vehicle (EV) batteries for short-term load-shifting is possible, but the large scale transition to EVs is yet to come. It is to cover this need that thermal storage in buildings can prove to be a key solution. As heating becomes more electricity based, a huge potential for energy buffering and flexible load can be released (Palensky and Dietrich, 2011). Approximately 40% of the total energy consumption is related to buildings, either used for heating or consumed by electrical appliances. Thermal energy can either be stored in the building structures, making it possible to shift load while keeping the indoor temperature within some limits (e.g. 19 to 21 °C), or stored in a thermal storage system, for example a hot water tank. This applies both to heating and cooling, or both at the same time as suggested by Blarke et al. (2012), who finds that thermal storage is a very cheap compared to electro-chemical or mechanical storage.

Apart from the hardware for enabling the use of buildings for energy storage, the methods for optimal operation are necessary to have. Modelling and forecasting of the energy flows in the building are needed, together with methods for modelling the energy systems related locally to the building. For example for forecasting of solar thermal power, if a solar collector is connected to the heating system. Finally, the operation of the heating system must be coupled to the operation surrounding energy system. This will most likely be carried out with energy markets, where a variable energy price will reward a flexible load. Today smart meters are being installed in many buildings and the high

resolution readings (10 to 15 minute values) of heat load will - combined with climate data - form the basis for modelling the building energy characteristics. Especially a proper modelling of the heat dynamics is crucial for this time resolution and provides the key for optimal use of the buildings for load-shifting. Here it is also emphasized, that another very important use of data from smart meters is for determination of the energy performance of buildings, as carried out by ENFOR (2010) and as with the methods presented in below. This will be very valuable for efficient energy refurbishment of buildings, which has a very important feature that must not be overseen: energy savings are far less volatile and to a large extent cheaper, compared to an equivalently increased renewable energy production.

In the thesis two methods for modelling related to building energy use are presented: First a method for forecasting the heat load of single-family houses and secondly a method for obtaining a detailed description of the heat dynamics of a building.

The method for forecasting of heat load is very well suited for optimized operation of a heating system, which include thermal storage in some dedicated medium, for example a hot water tank. As described previously, the objective of the PhD-project is to make methods to be used for optimal operation of a solar-electric combisystem for space and hot water heating in single family houses. The idea is to couple forecasts of: the electricity price, the energy production from the solar thermal collector, and the heat load of the building. The forecasts are then used as input to economic model predictive control scheme (MPC), which then provides an optimal operation plan for the system, i.e. how much electricity should be bought for heating at a given time. If the electricity can be bought at a variable price, then the storage capacity of the hot water tank can be used for lowering the running costs, by shifting the load to hours with cheap electricity. Hence the system can provide services to the grid, and thereby the short-term energy storage for an efficient integration of the large amount of wind power planned in Denmark. It is planned to have around 50% of electricity production in 2020.

The method, which can be used to obtain a detailed and accurate description of the heat dynamics of a building from data, is based on grey-box modelling. The models are formed by stochastic differential equations, which enables a combination of physical and data-driven modelling. The physical part for the model is formed by a description of the heat dynamics with differential equations, which includes the physical thermal properties and a model of how the significant heat transfers occur in the building. This allows for the estimation of parameters, which are directly physically interpretable. The identified model can be used for several purposes. Firstly, thermal properties, such as the UA-value of the building, provide valuable information, which can be used to identify buildings with

high potentials for energy refurbishment. Secondly, the knowledge of the heat dynamics can be used to optimize the operation of the heating system, while keeping the indoor temperature within some limits, as mentioned above, and for example minimize the operation costs, when the electricity can be bought at a variable price. Model predictive control can be used for carrying out the optimization using the model of the dynamics, as described by Šíroký et al. (2011) and Zong et al. (2011). This will enable the thermal mass of buildings to be used for load-shifting

In the following chapters the methods, which are presented in the included papers, are outlined and discussed. In Chapter 2 an overview is given of the methods relating to solar energy applications, which are presented in Paper A to E. In Chapter 3 an overview is given of the methods related to energy use in buildings, which are presented in Paper F to G. Chapter 4 finalizes the thesis with a general discussion and conclusion.

CHAPTER 2

Methods for solar energy applications

An overview of the presented methods, which can be used for the different purposes related to solar energy, is given in this chapter. First a statistical clear-sky model, which is used in several different contexts, is presented. For example in the method for correction of systematic errors in global radiation observations, which is outlined subsequently. This is followed by a section in which two methods for point forecasting and an approach for probabilistic forecasting of solar power are described. The chapter ends with a section on grey-box modelling of solar thermal collectors and finally a discussion is given.

2.1 Statistical clear-sky model

Clear-sky models, as described by Bird and Riordan (1984) and Rigollier et al. (2000), are primarily used for calculation of the global radiation under clear sky at some point in time and some location on the surface the earth. They are based on physical modelling of the radiative transfer of solar radiation through the atmosphere. One application of clear-sky models is for making a normalization of observations from sensors, which outputs are directly related to the solar radiation, for example the output of a pyranometer or a PV system. By normalization of the output it becomes evenly distributed throughout the day and thus

becomes a stationary process. The physical based clear-sky models take many effects into account, however they cannot account for effects related locally to the sensor, such as: bias in the sensor, shadows from trees, or reflections from a nearby water surface. In this thesis a statistical clear-sky model is presented, which can account for such effects. This is achieved by modelling the clear-sky output with a statistical model based solely on the data from the sensor. The models provide an estimate of the sensor output at clear-sky conditions for a given time. The estimated clear-sky output for the sensor is very well suited for the normalization and can be used for several purposes - in the thesis it is used for solar power forecasting and, together with a physical clear-sky model, for correction of typical errors in global radiation observations. The principles behind the statistical clear-sky model is the use of a locally weighted quantile regression model, where a quantile close to one is estimated for the given solar output. Depending on the application the statistical clear-sky model can be composed differently. The most general definition is with second order polynomial quantile regression

$$\hat{\beta}_t = \underset{\beta \in \mathbb{R}^5}{\operatorname{argmin}} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \rho_{\tau}(G_{t+24i+j} - (\beta_{0,t} + \beta_{1,t}i + \beta_{2,t}i^2 + \beta_{3,t}j + \beta_{4,t}j^2))K(i,j) \quad (2.1)$$

where $\{G_t, t = 1, \dots, N\}$ is a time series with hourly values (e.g. global radiation observations), $\rho_{\tau}(u) = u(\tau - I(u < 0))$ (see (Koenker, 2005) and (Koenker, 2011)), $\tau \in [0, \dots, 1]$ is the sample quantile to be estimated (which should be close to 1), $i \in \mathbb{N}$ is a counter of days, $j \in \mathbb{N}$ is a counter of hours, and $K(i, j)$ is a kernel function. The estimated clear-sky value at time t is then found as the intercept

$$G_t^{\text{cs}} = \beta_{0,t} \quad (2.2)$$

The weights are calculated with the two dimensional multiplicative Epanechnikov kernel function

$$K(i, j) = \begin{cases} \frac{3^2}{4^2} \left(1 - \left[\frac{|i|}{h_{\text{day}}}\right]^2\right) \left(1 - \left[\frac{|j|}{h_{\text{tod}}}\right]^2\right) & \text{for } \frac{|i|}{h_{\text{day}}} \geq 1 \wedge \frac{|j|}{h_{\text{tod}}} \geq 1 \\ 0 & \text{for } \frac{|i|}{h_{\text{day}}} < 1 \vee \frac{|j|}{h_{\text{tod}}} < 1 \end{cases} \quad (2.3)$$

where h_{day} is the bandwidth in the *day of year* dimension and h_{tod} is the bandwidth in the *time of day* dimension. The local weighting is thus carried out in the *day of year* and *time of day* dimensions, such that observations which are close in these two dimensions are given a higher weight in the estimation. The strong feature of using these dimensions is that they are directly related to both: the position of the sun and changing conditions over time. In Section A.3 (page 51) and Section B.3 (page 71) more detailed descriptions of the statistical clear-sky model are found, and in the following it is presented together with its applications.

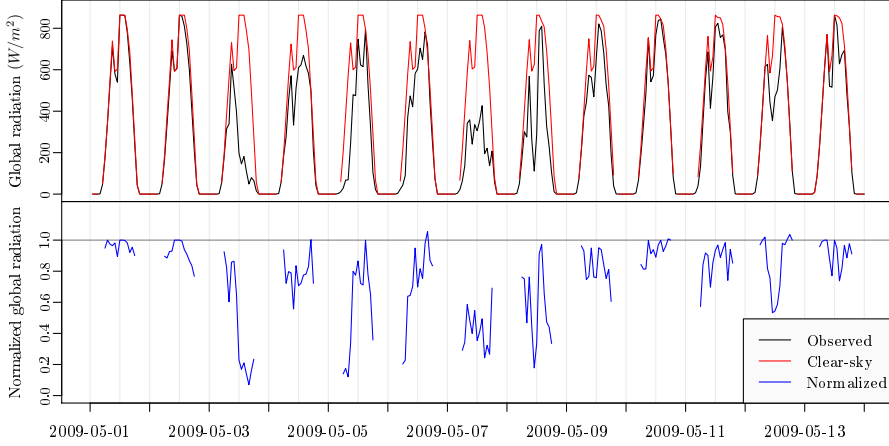


Figure 2.1: The upper plot is of the observed radiation and the estimated clear-sky radiation. The lower plot is of the normalized global radiation.

2.1.1 Correction of global radiation observations

Observations of global radiation is an important source of information, especially for solar energy applications. The observations are typically exposed to several types of systematic errors, for example: shadows from objects in the surrounding, tilt in levelling of the sensor, and processing errors (Younes et al., 2005). In Paper A a method for correction of such errors is presented. The correction is carried out using the statistical clear-sky model, for both the observations G_t and NWP of global radiation G_t^{nwp} in order to find the systematic deviance between them. The time series consist of hourly average values. Instead of the NWPs, a clear-sky model based on atmospheric physics can be used, e.g. (Bird and Riordan, 1984). The upper plot in Figure 2.1 shows the observed global radiation together with the estimated clear-sky radiation. It can be seen how the estimated clear-sky radiation follows the observed radiation on clear-sky days. Two types of errors are readily seen: there is a drop before noon and the observations are clipped at a maximum level. It can be seen that the errors are apparent in the observations and also described by the clear-sky model. The weather station, at which the observations were recorded, is located at a district heating plant. It is the plant chimney which shades and cause the drop, see the image on page 47. The lower plot in Figure 2.1 shows the normalized solar power, which is obtained by simple division of the observed radiation with the

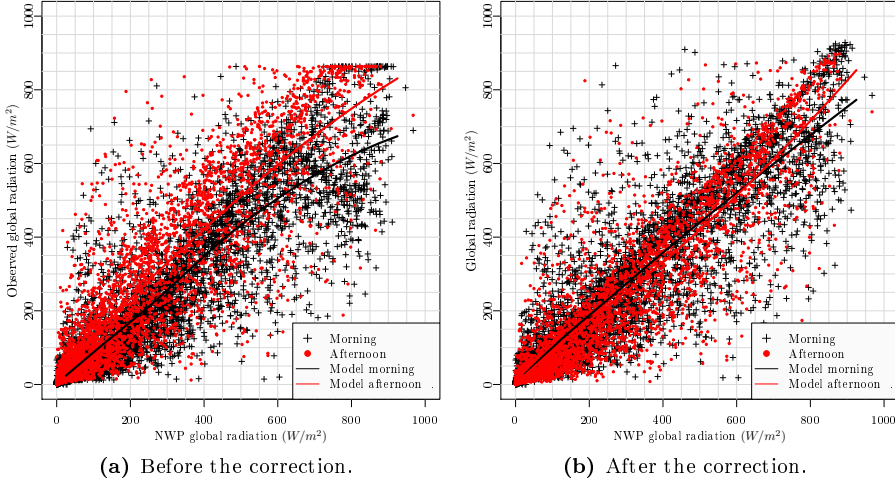


Figure 2.2: The values of the observations versus the NWPs before and after the correction. The morning values and the afternoon values are indicated by different symbols and colors. The two lines show a locally weighted least squares regression estimate of the relation between the variables in the morning and the afternoon.

estimated clear-sky radiation

$$\tau_t = \frac{G_t}{G_t^{cs}} \quad (2.4)$$

This process is clearly much less dependent on the time of day, hence it is a more stationary process, and are much less affected by shadow-drop before noon and saturation. The normalized process is multiplied with the estimated clear-sky radiation for the NWPs and thereby the corrected observations

$$G_t^{co} = G_t^{nwp,cs} \frac{G_t}{G_t^{cs}} \quad (2.5)$$

are obtained. It is furthermore noted that a tilt in the levelling of the sensor cause the observed level to be generally lower in the morning compared to the afternoon. The result of the correction can be seen by comparing the two scatter plots in Figure 2.2. The left scatter plot is of observations versus the NWPs before the correction, where both the shading and the tilt error can be seen. The right plot is a similar plot after the correction. It can be seen that the errors are mostly removed.

2.2 Solar power forecasting

In the following sections an overview of the solar power forecasting methods presented in the papers is given. The methods are based on modelling the power output past data consisting of past measurements and NWP. The forecast horizons are as long as provided by the NWPs, in presented applications they are complete up to 42 hours ahead at any time of day. Two basic approaches are presented: A two-stage method based on the statistical clear-sky model combined with a linear model, and a one-stage method based on a conditional parametric model. The two-stage approach is applied to forecasting of the total output of 21 PV-systems located in a small village in Denmark, it is described in Paper B. The one-stage approach is applied to forecasting of the output of both: a PV-system, as described Paper C, and to forecast the output of a solar thermal collector, as described in Paper D. Finally, an approach to probabilistic solar power forecasting is outlined in Section 2.2.3.

The obtained results from application of the forecasting methods are based on the data described in the papers, together with NWPs from the Hirlam-S05 (DMI, 2011), which are provided by the Danish Meteorological Institute.

2.2.1 Two-stage method based on the statistical clear-sky model

In paper B a forecasting method is described, where first the statistical clear-sky model is used to normalize the solar power

$$\tau_t = \frac{P_t}{\hat{P}_{cs,t}} \quad (2.6)$$

where P_t is the solar power, $\hat{P}_{cs,t}$ is the estimated clear sky solar power and τ_t is the normalized solar power. The NWPs of global radiation are similarly normalized to $\hat{\tau}_{t+k|t}^{nwp}$. This is carried out to obtain more stationary processes, such that the distribution is much less dependent on the day of year and to the time of day. The normalized processes is in the range of 0 to 1 and have approximately the same distribution at any time t . This is similar to the normalization of global radiation described in the previous section and illustrated in Figure 2.1, where the lower plot shows the normalized process. The normalization allows for of classical linear time series models (Box et al., 1976) to be used for forecasting. The best performing model is identified to the ARX model

$$\tau_{t+k} = m + a_1\tau_t + a_2\tau_{t-s(k)} + b_1\hat{\tau}_{t+k|t}^{nwp} + e_{t+k} \quad (2.7)$$

where $s(k)$ ensures that the latest diurnal component is used (see page 85). The same ARX model is fitted for each horizon, using a recursive least squares with forgetting scheme to achieve time adaptivity. Finally, the forecasts are transformed back again with the clear-sky model. The results show a root mean square error (RMSE) improvement of around 35 % over a persistence reference model.

2.2.2 Solar power forecasting with a conditional parametric model

An alternative to the statistical clear-sky model for compensating for the non-stationarity of the solar power is to use conditional parametric models, see (Nielsen et al., 2000) for more details about conditional parametric models. The same basic approach, as for the statistical clear-sky model, is taken by using a model, which is fitted locally in the *day of year* and *time of day* dimensions. Furthermore the conditional parametric models allows for the model to be non-linear in the inputs, for example as a non-linear function of the NWP of global radiation. For improving the forecasting performance for short horizons (up to three hours) the latest available observation is also added as input and the best model is obtained as

$$P_{t+k} = m + a(t_{\text{day}}, t_{\text{tod}}, P_t)P_t + b(t_{\text{day}}, t_{\text{tod}}, G_{t+k|t}^{\text{nwp}})G_{t+k|t}^{\text{nwp}} + e_{t+k} \quad (2.8)$$

where the coefficient function $a(t_{\text{day}}, t_{\text{tod}}, P_t)$ and $b(t_{\text{day}}, t_{\text{tod}}, G_{t+k|t}^{\text{nwp}})$ are non-linear functions, see Paper C for more details. For illustration of this, the plot in Figure 2.3 is used. It illustrates the 24 hour ahead forecasts of hourly solar power as a function of the NWP of global radiation for five days over the year. From these plots it can be seen how the forecasting function can be non-linear and how it changes conditional on the *day of year*. The results for forecasting of the power output from a PV system, as presented in the paper, shows a slight RMSE performance improvement compared to the two-stage method. The approach is also very well suited for forecasting of thermal power output from a solar collector, as presented in Paper D, where it is found that especially the possibilities for modelling non-linearity in the inputs are important.

2.2.3 Probabilistic solar power forecasting

Probabilistic forecasting is not presented in any of the papers in this thesis, however it is found that a natural next step forward is to expand the current

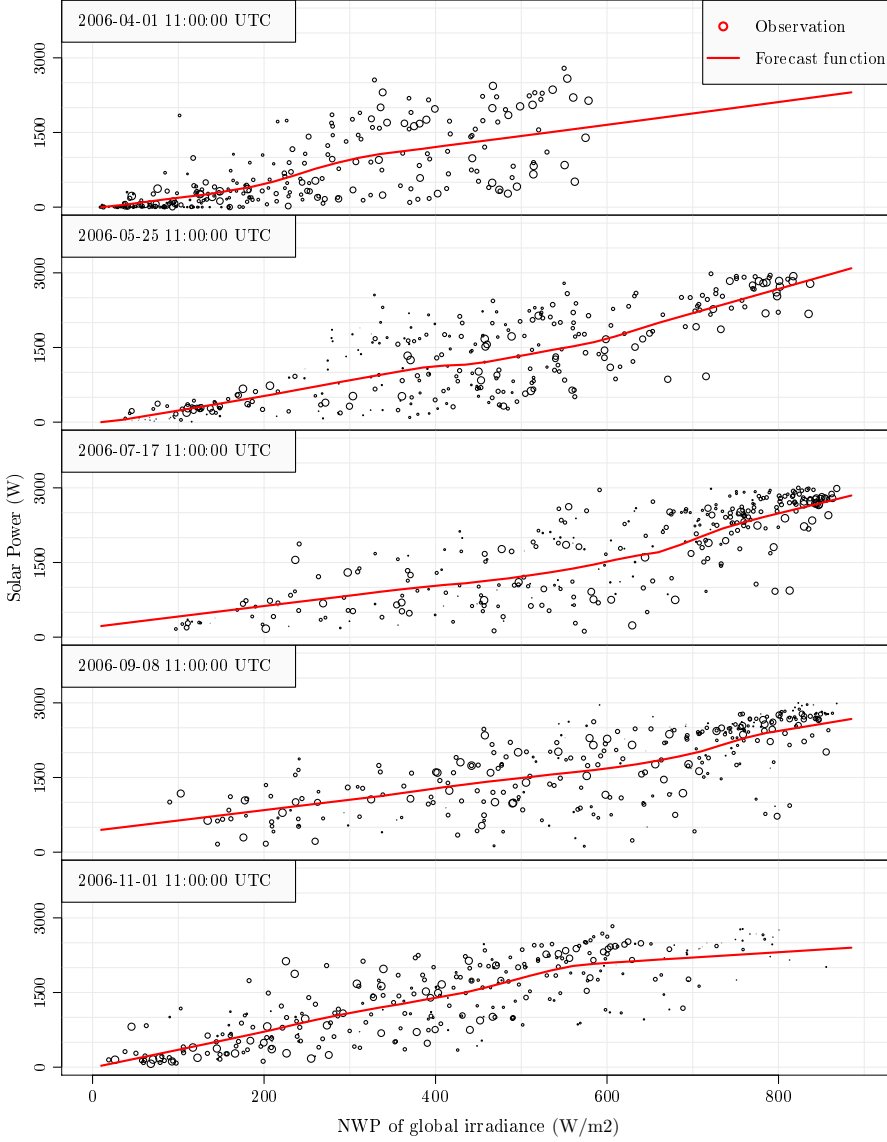


Figure 2.3: Example of the 24 hour ahead forecasted hourly solar power as a function of the NWP of global radiation for five days over the year. The data is presented in Paper C. The weighting of a point (only by $day(t, i)$ and $tod(t, i)$) is indicated by the size of its circle in the plot. A nearest neighbor weighting scheme is used.

point forecasts to probabilistic forecasts. Probabilistic solar power forecasting - which will be valuable for energy system operation, as described for wind power Pinson et al. (2007a) - means that at a given time ahead the probability distribution of solar power is forecasted instead of the single value. Many approaches to probabilistic forecasting exist, here an approach using quantile regression is used for illustration of some important aspects. One aspect is the relation between the normalized solar power - which can be considered equivalent to the clearness index - and the uncertainty on the forecast becomes apparent by considering normalized NWP's plotted versus normalized solar power, as described in Section B.5 (page 82). It is found that the uncertainty of the forecasts is dependent on the cloud cover, which is described with the clearness index, in such a way that forecasts for overcast or cloudless conditions have lower uncertainty, than forecasts of partly cloudy conditions. This is consistent with findings in other studies (Lorenz et al., 2007).

Probabilistic forecasting with quantile regression has been successfully applied for wind power forecasting by Bremnes (2004) and Nielsen et al. (2006). The suggested approach here for solar power is based on the same basic principles as the forecasting with a conditional parametric model, as presented in the previous section and in Paper C. The same data, consisting of observed solar power P_t and NWP's of global radiation $G_{t+k|t}^{\text{nwp}}$, is used, and the forecasting function is simply replaced with the local quantile regression model

$$\hat{\beta}_t = \underset{\beta \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^t \rho_{\tau} \left(P_i - (\beta_{0,t} + \beta_{1,t} G_{i|i-k}^{\text{nwp}}) \right) K_{3d}(t, i) \quad (2.9)$$

where i and t are denoting time, $\rho_{\tau}(u) = u(\tau - I_{\{u < 0\}})$ is the quantile regression objective function (see (Koenker, 2005)), and the 3-dimensional multiplicative kernel function

$$K_{3d}(t, i) = K(\text{day}(t, i)) \cdot K(\text{tod}(t, i)) \cdot K(G_{t|t-k}^{\text{nwp}} - G_{i|i-k}^{\text{nwp}}) \quad (2.10)$$

where $\text{day}(t, i)$ is the distance from t to i in days, $\text{tod}(t, i)$ is the time of day distance, and the Epanechnikov kernel function

$$K(\Delta) = \begin{cases} \frac{3}{4} \left(1 - \left[\frac{|\Delta|}{h} \right]^2 \right) & \text{for } \frac{|\Delta|}{h} \geq 1 \\ 0 & \text{for } \frac{|\Delta|}{h} < 1 \end{cases} \quad (2.11)$$

with bandwidth h is used for calculating the weights in each dimension. The following bandwidths, which was set from visual inspection of the results, are used in the three weighting dimensions:

- $h_{\text{day}} = 150$ is bandwidth in days

- $h_{\text{tod}} = 2$ is bandwidth in hours
- $h_{G_{\text{nwp}}} = nn(t, i, 300)$ is bandwidth in W/m^2 calculated as the smallest bandwidth where the 300 nearest neighbors are included.

The quantile regression implementation in the **R** (R Development Core Team, 2011) package '`quantreg`' is used for carrying out the calculations.

The plots shown in Figure 2.4 is of the estimated quantiles conditioned on the NWP of global radiation, for a 24 hour horizon. It is the 5%, 10%, ..., 90%, 95% τ quantiles. A clear dependency on the level of the global radiation NWP is seen and the same pattern as described above is found: the uncertainty is higher for values in the middle of the range of forecasted global radiation.

It is evident that more research in probabilistic forecasting of solar power is needed, for example a proper evaluation should be carried out and an optimization of the bandwidths. From the very coarse outline of the presented approach a few conclusions are drawn:

- Evaluation of probabilistic solar power forecasts is needed for assessment of the forecast quality. Solar forecasting is highly related to wind power forecasting and the evaluation can be based on methods for evaluation of probabilistic wind power forecasts, for example the framework proposed by Pinson et al. (2007b).
- It is clearly seen that the distribution of solar power conditional on the NWP of global radiation is highly skewed for low and high levels of the NWP. Hence it can be questioned how a proper evaluation of point forecasts should be carried out, since the optimality of the least squares criteria is mostly based on an assumption of normal distributed errors.
- The value of energy forecasts for operation increase as more useful information for system operators is available. In that respect probabilistic solar power forecasting can prove to be very valuable, since the distribution of the error is highly dependent on the inputs and thus contain quite a lot non-trivial information.
- Studies have shown that valuable information about the uncertainty on wind power forecasts can be gained from ensemble forecasting. Ensemble forecasts calculated by the meteorologists can be used as input to model the uncertainty in energy forecasts, for example as described by Nielsen et al. (2006) for probabilistic wind power forecasting. The same approach can be applied for solar power forecasting.

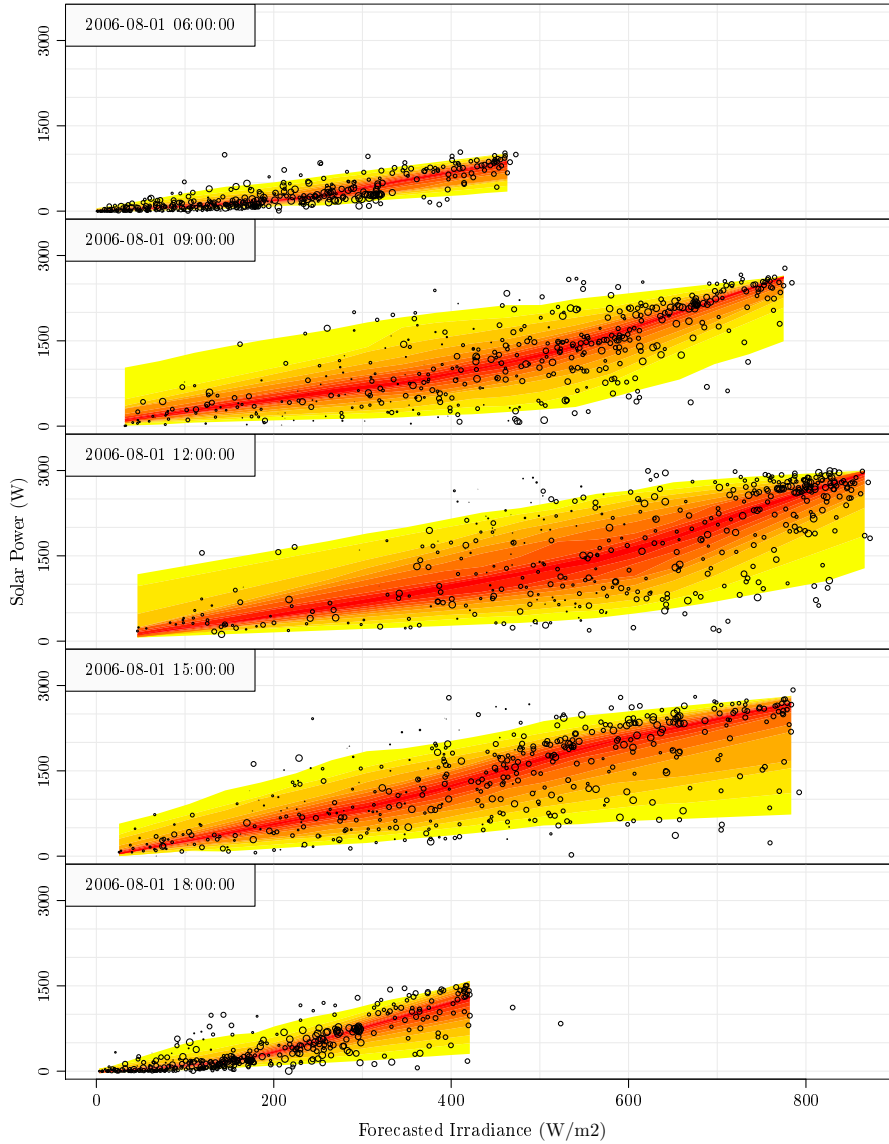


Figure 2.4: Example of probabilistic solar power forecasts for a $k = 24$ hour horizon at five different hours over the day. Local quantile regression is applied to estimate the 5%, 10%, ..., 90%, 95% τ quantiles. The weighting of a point (only by $day(t, i)$ and $tod(t, i)$) is indicated by the size of its circle in the plot.

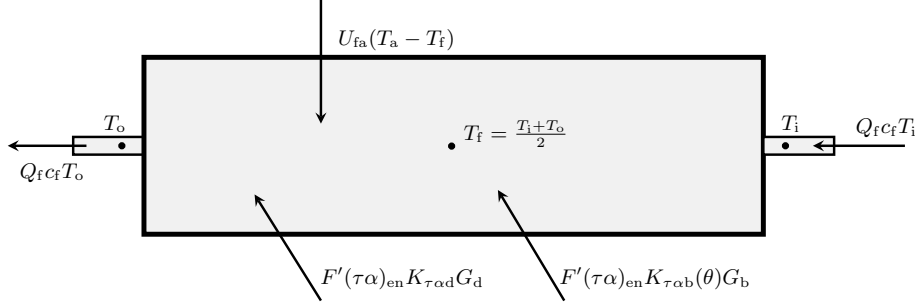


Figure 2.5: Diagram of the single state grey-box model illustrating all the energy flows included in the model.

2.3 Modelling the heat dynamics of solar thermal collectors

In Paper E models for the heat dynamics of solar thermal collectors are presented. The heat dynamics are described with grey-box models. A grey-box model is established using a combination of prior physical knowledge and statistics Kristensen et al. (2004). The prior physical knowledge is formulated by a set of non-linear stochastic differential equations (SDEs). The equations describe a lumped model of the heat dynamics of the system. Models from Perers (1997) of solar thermal collectors based on prior physical knowledge are used to derive the SDE which forms the system equation

$$dT_o = \left(F'U_0(T_a - T_f) + c_f Q_f(T_i - T_o) + F'(\tau\alpha)_{en}K_{\tau\alpha b}(\theta)G_b + F'(\tau\alpha)_{en}K_{\tau\alpha d}G_d \right) \frac{2}{(mC)_e} dt + \sigma d\omega \quad (2.12)$$

of the simplest applied grey-box model and it consists of a single state T_o which is the outlet temperature of the collector, [°C]. In the diagram in Figure 2.5 the energy flows are illustrated.

The following are parameters in the model

$F'U_0$	Heat loss coefficient at $(T_a - T_f) = 0$, $[W/(m^2K)]$.
$F'(\tau\alpha)_{en}$	Zero loss efficiency for direct radiation at normal incidence
$K_{\tau\alpha b}(\theta)$	Incidence angle modifier for direct radiation. It has a single parameter b_0 which is estimated
$K_{\tau\alpha d}$	Incidence angle modifier for diffuse radiation
$(mC)_e$	Effective thermal capacitance including piping for the collector, $[J/(m^2K)]$.
σ^2	is the incremental variance of the Wiener process $\{\omega\}$

The following are inputs which are measured (or derived directly from measurements)

T_a	Ambient temperature, $[^\circ C]$.
T_f	Average temperature of the collector fluid, $[^\circ C]$.
Q_f	Flow of the fluid per square meter of collector, $[l/(sm^2)]$.
T_i	Temperature of the inlet to the collector, $[^\circ C]$.
G_b	Direct radiation onto the collector plane, $[W/m^2]$.
G_d	Diffuse radiation onto the collector plane, $[W/m^2]$.

and finally c_f which is the specific heat capacity of the fluid, $[J/(l K)]$.

The data-driven part of the grey-box model is formed by the discrete time observation equation

$$Y_k = T_{ok} + e_k \quad (2.13)$$

where k is the point in time t_k of a measurement, Y_k is the measured outlet temperature, and e_k is the measurement error, which is assumed to be a Gaussian white noise process with variance σ^2 . The parameters are estimated with maximum likelihood techniques as described in Section E.2.1 on page 130.

A suitable model is identified for a flat plate collector using 30 second average values measured at a test setup at the building department of DTU. It is shown how the models can provide detailed knowledge about the performance of the collector with data from a single day. A forward model selection approach is used, where first a simple model is applied and stepwise expanded. The final model includes two lumped parts: one representing the fluid and one representing the metal and frame of the collector. Each of the two parts are lumped further into three compartments, resulting in the final model having six state variables, see Section E.5.2 on page 136.

A strong feature of using the applied grey-box modelling, is that accurate physical knowledge can be obtained with a minimum of testing time. The results are evaluated and discussed both from a statistical and physical perspective. As a reference the equivalent parameter estimates found with current ISO-standardized multiple linear regression models are used. The results show that parameters can be estimated very accurately with 30 second average values from a single day. It is an prerequisite that the input, especially the ambient temperature and the global radiation is uncorrelated. For the presented study this was achieved by using data from a day, where clouds modulated the radiation with an "on-off" sequence. This should be replaced with a shading device - simply a foil that can be automatically rolled up and down - for enabling an "on-off" testing sequence for the radiation onto the collector. The models can be used in several contexts, both for performance testing and for operational applications.

2.4 Discussion

Two methods for solar power forecasting based on NWP have been presented, together with a method for correction of measured global radiation, and an outline of a probabilistic solar power forecasting method. The recurrent principle, on which all of the methods is based, is the use of non-parametric local regression techniques to model measured solar output, by using a kernel weighting function in dimensions of *day of year* and *time of day*. The advantage of using such models is that all types of systematic effects embedded in the solar output are included in model, especially effects which are local to the sensor and therefore very complex to model with a physical model, e.g. shadowing from trees. Regarding the statistical clear-sky model, a local fitting in the dimensions of the *sun azimuth* and *sun elevation* has been considered in (Bacher, 2008), where it was found that it leads to decreased accuracy for the considered applications. This is due to the importance of the model being local in time, for example at the positions of the sun where trees are shading, then it makes a huge difference if there are leaves on the trees or not.

Regarding the solar forecasting methods presented it is clear that further work should be addressed to enhance the models. For example regime switching for using different models depending on the weather type and using multiple NWPs as inputs. This has proved to increase performance both for wind power forecasts and solar forecasts. For example Schmelter et al. (2011) presents a method for combining forecasts depending on a weather classification for improving solar power forecast performance, especially for handling special weather conditions like fog and snow.

CHAPTER 3

Methods for building energy applications

The methods outlined in this chapter can be used to deal with important aspects for optimized energy use in buildings, especially for enabling buildings as a key player in smart grids and for energy performance improvement of the building stock. The method for forecasting of the heat load presented in Paper F is outlined, together with the procedure for identification of a suitable model for the heat dynamics of buildings presented in Paper G. The data used consist of a combination of: measurements and NWP of climate variables, heat load, and indoor temperature.

3.1 Heat load forecasting

A method for forecasting the heat load in single family houses is presented in Paper F. Adaptive linear time series models are applied to forecast the heat load

$$Q_{t+k} = Q_{t+k|t}^{\text{ambient}} + Q_{t+k|t}^{\text{diurnal}} + Q_{t+k|t}^{\text{sun}} + Q_{t+k|t}^{\text{wind}} + e_{t+k} \quad (3.1)$$

where the terms on the right side represent heat gains caused by several mechanisms together with an error. The inputs to the model are formed from mea-

surements from a local weather station and numerical weather predictions. The following three signals are available

- $\hat{T}_{t+k|t}^a$ the ambient temperature
- $\hat{G}_{t+k|t}$ the global radiation
- $\hat{W}_{t+k|t}^s$ the wind speed

The heat dynamics of the building are described with linear rational transfer functions from the climate variables to the heat load. The heat gain to the ambient (i.e. negative in the winter period) is described by

$$Q_{t+k|t}^{\text{ambient}} = \alpha_{\text{ia}} + \alpha_a H_a(q) \hat{T}_{t+k|t}^a \quad (3.2)$$

where α_{ia} is representing a constant indoor temperature (which is not available and therefore modelled as an intercept) and the first order low-pass filter with unity DC-gain

$$H_a(q) = \frac{1 - a_{T_a}}{1 - a_{T_a} q^{-1}} \quad (3.3)$$

is applied and where q^{-1} is the backward shift operator ($q^{-1}x_t = x_{t-1}$) (see (Madsen, 2007)) and $a_{T_a} \in [0, 1]$ is a coefficient, which is equivalent to the time constant for the part of the building affected by changes in ambient temperature. This is equivalent to describing the heat dynamics of the building with a lumped RC-model having a single heat capacity for the interior of the building and a single thermal resistance through the building envelope. The diurnal curve heat gain describes diurnal patterns caused by systematic user behavior, for example a nightly setback. It is modelled as a harmonic function using a Fourier series

$$\mu(t_{\text{tod}}, \alpha_{\text{diu}}) = \sum_{i=1}^{n_{\text{har}}} \alpha_{i,1}^{\text{diu}} \sin\left(\frac{t_{\text{tod}} i \pi}{12}\right) + \alpha_{i,2}^{\text{diu}} \cos\left(\frac{t_{\text{tod}} i \pi}{12}\right) \quad (3.4)$$

where t_{tod} is the time of day in hours at time t and n_{har} is the number of harmonics included in the Fourier series. The heat gains from solar radiation and wind are modelled as

$$Q_{t+k|t}^{\text{ambient}} = \alpha_g H_g(q) \hat{G}_{t+k|t} \quad (3.5)$$

and

$$Q_{t+k|t}^{\text{wind}} = \alpha_w H_w(q) \hat{W}_{t+k|t}^s \quad (3.6)$$

where the low-pass filters $H_g(q)$ is similar to the filter for ambient temperature (Equation (3.3)), but fitted independently such that the coefficient describing

the dynamic response of the building is different for each of the inputs. Several extensions to this model is also applied as described in the paper. In order to achieve time adaptivity of the model, it is fitted with a recursive least squares scheme, where the past data is down-weighted depending on a forgetting factor. Note that each of the coefficients could have been denoted with a t to indicate that they change as a function of time. Finally, in a second stage a simple AR(1) model is applied to remove the last correlation in the residuals.

To identify a suitable forecasting model measurements of the hourly heat load for sixteen houses, which are typical Danish single family houses, are used. The houses are connected to district heating with a heat exchanger and have radiator heating. In order to lower the signal-to-noise ratio for effects related to the climate conditions, the heat load signals are preprocessed by filtering out the peaks from water heating. Thereafter a thorough model identification is carried out to find a forecasting model, which is suitable for all the houses. The following parameters are tuned for each house separately: transfer function coefficients (they are equivalent to time constants in the building and describe how fast the building respond to changes in the climate variables), harmonics in the diurnal curve, and the optimal time adaptivity. The first step in the model identification is to fit a simple model consisting only of a constant heat gain and a diurnal curve - i.e. this model does not include any climate variables, however it can follow the slow changes in climate due to the adaptive modelling scheme. This model is then expanded in steps, where inputs are included in different ways in a forward selection approach. The final result is a model, which is suitable for forecasting the heat load for each house, without requiring any specific knowledge about the building, apart from the heat load measurements. It can be used solely with NWP as input, but the addition of local measurements improves forecasting performance for short horizons. It is shown that the forecasting residuals are close to white-noise and thereby that the information embedded in the inputs are very well utilized.

The forecasting performance measured with the mean absolute error as a function of the forecasting horizon is plotted for each of the sixteen houses in Figure 3.1. Clearly, a quite large difference is found between the houses, especially the forecasting performance is poor for House 8. Analysing the forecasts for House 8 it is found that the heat load signal has some oscillations with a period around 6 hours, which are not possible to forecast and which are most likely caused by a poorly tuned thermostatic control. A further thorough analysis of the forecast for all the houses indicates that the solar radiation part of the model pose a challenge. Improvements could possibly be achieved by using some information about the buildings, such as the azimuth angles of the building walls etc., or by using an off-line model for learning how the solar energy gain of the building depends on the position of the sun, possibly as a function of *day of year* and *time of day*. Furthermore is seen how the error in the global radiation NWPs

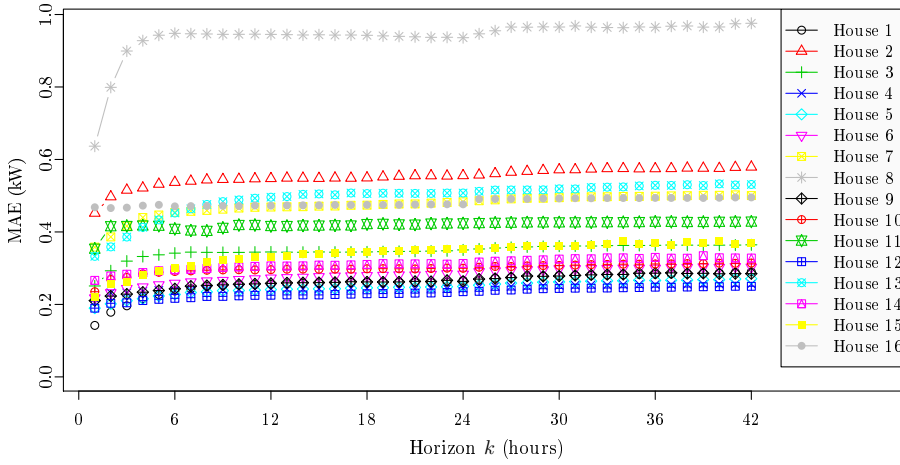


Figure 3.1: Mean absolute error for the hourly forecasted average values as a function of the forecasting horizon for each of the sixteen houses.

for longer horizons, i.e. an error in the input to the model, result in errors in the heat load forecasts. Furthermore, it is seen how effects related to unpredictable behavior of the residents, for example ventilation from opening of windows, appear in the heat loads.

3.2 Models for the heat dynamics of buildings

In Paper G models are presented, which can provide detailed knowledge of the heat dynamics of a building from measurements of: heat load, indoor temperature, ambient temperature, and global radiation. The focus of the paper is a procedure for selection of the most suitable model. The models are grey-box models, which are based on a combination of physical and data-driven modelling. They are based on stochastic differential equations, which allows for extensive modelling of dynamical systems and estimation of parameters which are directly physically interpretable, for example the heat capacity of the building and the UA-value of the building envelope. A description of heat transfer mechanisms which forms the basis for the physical part of the grey-box models can be found in (Bacher et al., 2010, Chapter 3).

The applied grey-box models are stochastic linear state space models, which are formed by a continuous time system equation and a discrete time measurement

equation. The system equation can be formulated in matrix form

$$d\mathbf{T} = \mathbf{A}\mathbf{T}dt + \mathbf{B}\mathbf{U}dt + \boldsymbol{\sigma}d\boldsymbol{\omega}(t) \quad (3.7)$$

where \mathbf{T} is a vector of state variables (typically both measured and unmeasured temperatures), \mathbf{U} is a vector of inputs, and $\boldsymbol{\omega}(t)$ is a Wiener process, which is a stochastic process with independent normal distributed increments. The matrices \mathbf{A} defines how the current state affects the dynamics and \mathbf{B} defines how input enters the system, and $\boldsymbol{\sigma}^2$ is the scaling of the linear in time growing variance of the increments of the Wiener process. The discrete measurement equation can be formulated as

$$\mathbf{Y}_{t_k} = \mathbf{C}\mathbf{T}_{t_k} + \mathbf{D}\mathbf{U}_{t_k} + \mathbf{e}_{t_k} \quad (3.8)$$

where t_k are the equidistant time points on which the output and inputs are measured, \mathbf{Y}_{t_k} is the measured output (typically temperatures and/or heat flows, for example the interior temperature), \mathbf{e}_{t_k} is the measurement error. It is assumed that \mathbf{e}_{t_k} is normal distributed white noise with zero mean and variance \mathbf{R}_e . Furthermore it is assumed that \mathbf{e}_{t_k} and $\boldsymbol{\omega}(t)$ are mutually uncorrelated. \mathbf{C} and \mathbf{D} defines how the measured states are influenced by the state and input respectively.

A lumped parameter model is used to describe the heat dynamics of the building as exemplified in the following, where a two state model is defined. One state variable is describing the interior temperature T_i and one is representing the temperature of the building envelope T_e . The first-order dynamics are represented by the stochastic differential equations

$$dT_i = \frac{1}{R_{ie}C_i}(T_e - T_i)dt + \frac{1}{C_i}\Phi_h dt + \frac{1}{C_i}A_w\Phi_s dt + \sigma_i d\omega_i \quad (3.9)$$

$$dT_e = \frac{1}{R_{ie}C_e}(T_i - T_e)dt + \frac{1}{R_{ea}C_e}(T_a - T_e)dt + \sigma_e d\omega_e \quad (3.10)$$

where t is the time, R_{ie} is the thermal resistance between the interior and the building envelope, R_{ea} is the thermal resistance between the building envelope and the ambient air, C_i is the heat capacity of the interior, C_e is the heat capacity of the building envelope, Φ_h is the energy flux from the heating system, A_w is the effective window area, Φ_s is the energy flux from solar radiation, T_a is the ambient air temperature, $\{\omega_{i,t}\}$ and $\{\omega_{e,t}\}$ are standard Wiener processes, and σ_i^2 and σ_e^2 are the scaling of the linear in time growing incremental variances of the Wiener processes. Note that the equations can easily be written into the matrix form defined in Equation (3.7). The model can be represented with the RC-network depicted in Figure 3.2, where the model is divided into different parts to show the corresponding parts of the building.

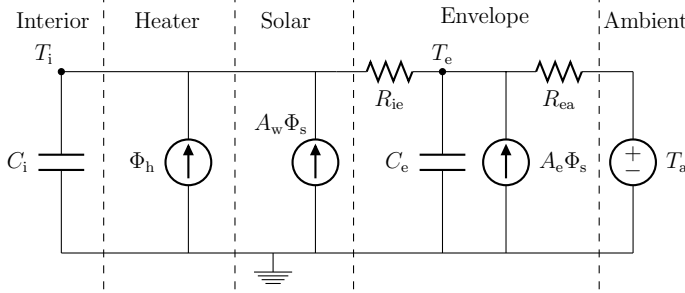


Figure 3.2: RC-network of the model described by Equation (3.9) and (3.10). The model is divided into different parts indicating the corresponding part of the building.

The interior temperature is measured and the discrete time measurement equation is

$$Y_{t_k} = T_{i,t_k} + e_{t_k} \quad (3.11)$$

where t_k is the time point of a measurement, Y_{t_k} is the measured interior temperature and e_{t_k} is the measurement error, which is assumed to be a Gaussian white noise process with variance σ^2 .

The presented procedure for identification is based on a forward model selection approach, where likelihood ratio-tests are used for selection of a grey-box model with a suitable complexity. This is important since, on the one hand, a too simple model will be biased and not model the dynamics with the required level of detail, and on the other hand, a too complex model will be over-parametrized leading to degraded performance and incorrect parameter estimates. The procedure is based on a forward selection approach, where a simple model is extended until no significant increase in likelihood is found.

In the paper the procedure is applied to identify a model for the heat dynamics of a single-storey 120 m² building. The building is heated with a 6 days test sequence and five minutes average values are used for the modelling. An image and floor plan of the building can be seen in the paper (page 184). First the very simple model illustrated with its RC-diagram in Figure G.7 (page 191) is fitted. Then an expansion of this is sought by adding different model parts, one at a time, and the expanded model with the highest increase in likelihood is selected. The expansion step is repeated iteratively until likelihood-ratio tests shows that no more significant increase in likelihood is achieved by any expansion. The final selected model is illustrated by its RC-diagram in Figure G.8 (page 193). The one-step prediction residuals for the models selected in each step are analysed with time series plots, the auto-correlation function and the cumulated periodogram. These can be seen in Figures G.9, G.10 and G.11. The

identified model can be applied for several purposes. The estimated parameters gives important information of the thermal performance of the building. The values can be found in Table G.3. The UA-value of the building envelope, which can be calculated directly from the parameters, indicates how well the building is insulated, see Table G.4. Furthermore accurate knowledge of the heat capacity of the building is obtained and thereby the necessary knowledge for using the model as basis for control is provided.

Due to the rather uncomfortable course of the indoor temperature the modelling method using the used test sequence is not directly applicable in residential buildings. However it can be applied in periods, where the residents are away from the house in some days or for example at night time and during the day. In office buildings it can be run during the weekends. More work is needed for applying the procedure with other buildings in order to determine the minimum requirements for a desired level of estimation accuracy. For MPC methods for optimal load-shifting control of heating systems, in which the models are used, for example Pedersen et al. (2011) and Zong et al. (2011) both suggest that the indoor temperature is to be kept within a given range, for example between 19 to 21 °C in some periods and within 18 to 22 °C in others. Identifying a model for control allowing a indoor temperature variation in this range is probably a feasible task.

A strong feature is the capabilities of modelling the uncertainty in SDE models, which can be important for including residents behavior in the models. For example it can be modelled such that it depends of the time of day, in order to describe an increased uncertainty when the residents are home.

Finally, it is mentioned that the presented grey-box modelling procedure is very well suited for effective modelling in relation to effective performance testing of building components, e.g. in a test cell, as described in (Madsen and Holst, 1995).

3.3 Discussion

Considering the variety in the building stock and the highly diversified use of buildings, it is clear that much more research in modelling of heat use in buildings, based on data from smart meters, is needed. As seen in the data used in Paper F even for the sixteen single-family houses the characteristics of the heat load signals are very different, see for example the plot on page 171. However, even if the models as presented in the two papers are applied for a relatively specific type of buildings, the underlying principles can be applied for

modelling dynamic systems in many respects, including many types of buildings.

The need for details and accuracy for efficient operation of energy systems related to buildings depends on the application and the availability of data. A statistical approach, which eventually enables nearly all available information in data to be modelled, can save costs for hardware, e.g. sensors. The level of forecast uncertainty will naturally be related to the operational performance of the energy system, but it is a (economical) trade-off between the need for information from sensors (how many sensors etc. that are needed) and the performance gain, which finally also maps to an economical value of operating the total energy system, since more expensive energy backup capacity is needed as the forecasting accuracy decrease. Determining the cost-optimal need for sensors and accuracy is a non-trivial task, but a very valuable question to be able to answer.

Modelling the uncertainty for probabilistic forecasting presents another challenge, since the dynamics of the system needs to be included. For example the errors from the NWP forecasts of solar radiation can have a significant and one-sided effect on the forecast uncertainty, and thus the dynamics needs to be taken into account for a proper modelling.

CHAPTER 4

Discussion and conclusion

The presented methods have been discussed separately, hence in this section a general discussion of the presented work is given.

4.1 General discussion

The methods are closely related by the data-driven approach using statistical times analysis techniques for modelling. They form a basis for a range of operational applications, which are needed as the amount of data acquired from the energy system is increasing rapidly, for example with the widespread installation of smart meters in buildings. The data will be used for providing the needed on-line modelling of the system, where knowledge of the state and the dynamics in different part of the system is vital for an efficient operation. It is widely recognized that the operation of the power grid will be carried with markets, where a variable electricity price is available on different timescales. This will reward consumers which can provide a flexible load. Both for the operator of such markets and the participants, e.g. a consumer offering flexible heat load services in buildings, forecasting and methods for optimal operation are prerequisites for enabling the needed flexible load. Furthermore it is noted that it is very important to avoid sub-optimization of interdependent energy systems, for example as for operation of the considered solar/electric heating system, where

it is very important that the operation of the solar heating and the electrical heating is jointly optimized. Hence the forecasting and modelling methods, as the ones presented, should be combined in order to achieve an efficient operation of smart grid enabling technologies.

The required accuracy for the methods is a very relevant, but non-trivial question, to consider. It is known from statistical modelling theory that performance of a model is decreased if it is too simple or too complex. Therefore models should only comprise the needed complexity and here the statistical techniques for optimization of model complexity are vital. They are for example demonstrated in Paper G with the procedure for model identification based on likelihood-ratio tests. A forward selection approach for model identification is used for most of the presented modelling, in order to optimize the model complexity. Roughly said this simply implies starting with a simple model and trying different expansions until no performance improvement is found. To some extent this will require some manual interaction, but most likely the methods can be applied automatically for the majority of cases, in a combination with an automatic selection of cases which require further manual interaction. Certainly the computational resources needed is also an issue when considering operation of many thousand units. The presented methods are very suited for on-line operation, especially the recursive schemes for forecasting are very computationally efficient. Generally it is suggested to run the computations on a central server. This will allow for easier management and more efficient computation, for example some calculations can be carried only once, opposed to stand-alone implementation.

Modelling of the forecast uncertainties is a really important step for power grid operation, as the share of renewables in the power mix is increasing, the forecasting errors will have a greater impact on the operation of the power system. Hence the value of modelling the uncertainties, and have probabilistic forecasting for all different types of power generation and load in the system, will increase. Considering the sources of errors there is a distinction between modelling errors, e.g. errors in the meteorological forecasts, and errors caused by for example user behavior. Errors from meteorological forecasts will mostly be affecting forecasts of wind and solar power generation, but also for the load, as especially the thermal part is included. Errors from user behavior will be affecting the load, but will, when considering the total load be smoothed out due to averaging over many individuals. However for operation of for example the heating system for a single house, the user behavior is very important to take into account.

4.2 Conclusion

Methods for optimized operation of energy systems with substantial amounts of renewable and volatile energy production are presented in the thesis. The methods are very well suited for smart grid applications. The focus is on electrical and thermal solar energy, and on heating in buildings.

The methods are mainly based on data-driven statistical models, which are combined with prior-physical knowledge where appropriate. The modelling carried out takes important aspects into account, such as: dynamical effects, non-stationarity, time adaptivity, non-linearity, user behavior, physical relations, and computationally efficient implementation. In the presentation of the methods, focus is on model identification and performance evaluation, which are carried out with statistical time series analysis techniques. The presented methods are almost directly applicable for operational use without much further development. Many perspectives and ideas for further work on the methods and in the respective fields are given.

A statistical clear-sky model is presented. It can be used to estimate the clear-sky output of a system, where the output is directly dependent on solar radiation at the surface of the earth. It is based solely on observations of the output and is very useful for removing non-stationarity in the observations. It also includes the effects occurring in the system and caused by the local surroundings, for example tilt in the levelling of the sensor (e.g. a pyranometer or a PV panel) or shading from objects in the surroundings. Two applications, where the statistical clear-sky model is used, are presented in the thesis: a two-stage method for solar power forecasting for PV-systems and a method for correction of errors in global radiation observations.

A second approach to solar power forecasting is presented. It is based on a conditional parametric model, which is conditional on the *day of year* and on the *time of day*. It is demonstrated how the method is well suited for forecasting solar power both for PV and solar thermal systems, and furthermore it is outlined how the approach can be used for probabilistic solar power forecasting.

Grey-box models for the heat dynamics of solar thermal collectors are presented. The models can for example be used to obtain accurate estimates of the energy performance parameters based on a very short testing period.

A method for forecasting the heat load for single family houses is presented. A forecasting model is found on the basis of heat load measurements from sixteen houses and local climate measurements combined with weather forecasts. The model is tuned to describe the heat dynamics of the building and the resident

behavior for each individual house. It is shown that practically all information embedded in the inputs are modelled.

Finally, a procedure for identification of a suitable model of the heat dynamics of a building is presented. The applied grey-box models are based on stochastic differential equations, which provide a detailed description of the heat dynamics based on prior physical knowledge combined with data-driven modelling. The models can be used as basis for model predictive control to enable shifting of the heat load, and for obtaining detailed knowledge of the energy performance of buildings.

Bibliography

- P. Bacher. Short-term solar power forecasting. Master's thesis, Technical University of Denmark, 2008. IMM-M.Sc.-2008-13.
- P. Bacher, A. Thavlov, and H. Madsen. Models for energy performance analysis : Financed by the danish electricity saving trust. Technical report, DTU Informatics, Building 321, Kgs. Lyngby, 2010.
- R. Bird and C. Riordan. Simple solar spectral model for direct and diffuse irradiance on horizontal and tilted planes at the earth's surface for cloudless atmospheres. Technical report, Solar Energy Research Inst., Golden, CO (USA), 1984.
- M. B. Blarke, K. Yazawa, A. Shakouri, and C. Carmo. Thermal battery with co2 compression heat pump: Techno-economic optimization of a high-efficiency smart grid option for buildings. *Energy and Buildings*, pages –, 2012. ISSN 0378-7788. doi: 10.1016/j.enbuild.2012.03.029.
- G. Box, G. Jenkins, and G. Reinsel. *Time series analysis*. Holden-day San Francisco, 1976.
- BP. Statistical review of world energy 2011, 2011. URL www.bp.com/statisticalreview.
- J. B. Bremnes. Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, 7(1):47–54, 2004. ISSN 1099-1824. doi: 10.1002/we.107.
- C. Breyer, S. Rieke, M. Sterner, and J. Schmid. Hybrid pv-wind-renewable methane power plants - a potential cornerstone of global energy supply. In *Proceedings of ISES Solar World Conference 2011*, 2011.

- Danish Commission on Climate Change Policy. Green energy – the road to a danish energy system without fossil fuels, September 2010.
- DMI. Danish Meteorological Institute, DMI-HIRLAM-S05, 2011. URL http://www.dmi.dk/eng/index/research_and_development/dmi-hirnam-2009.htm.
- ENFOR. Analysis of energy consumption in single family houses. Technical report, , 2010. URL <http://www.enfor.dk/pub/03EKS0009A002-A.pdf>.
- J. Fonseca Júnior, T. Oozeki, T. Takashima, and K. Ogimoto. Analysis of the use of support vector regression and neural networks to forecast insolation for 25 locations in japan. In *Proceedings of ISES Solar World Conference 2011*, 2011.
- V. Fthenakis, J. E. Mason, and K. Zweibel. The technical, geographical, and economic feasibility for solar energy to supply the energy needs of the us. *Energy Policy*, 37(2):387 – 399, 2009. ISSN 0301-4215. doi: 10.1016/j.enpol.2008.08.011.
- R. Halvgaard, N. K. Poulsen, H. Madsen, and J. B. Jørgensen. Economic Model Predictive Control for Building Climate Control in a Smart Grid. In *2012 IEEE PES Innovative Smart Grid Technologies (ISGT)*, volume 9781457721588, page 2012ISGT0195. IEEE, 2012.
- D. Heide, L. von Bremen, M. Greiner, C. Hoffmann, M. Speckmann, and S. Bofinger. Seasonal optimal mix of wind and solar power in a future, highly renewable europe. *Renewable Energy*, 35(11):2483 – 2489, 2010. ISSN 0960-1481. doi: 10.1016/j.renene.2010.03.012.
- IEA - The Solar Heating and Cooling Programme. Solar heat worldwide, 2010. URL <http://www.iea-shc.org>.
- IEA, International Energy Agency. *Solar Energy Perspectives - Executive Summary*. OECD Publishing, 2011a. doi: <http://dx.doi.org/10.1787/9789264124585-en>.
- IEA, International Energy Agency. World energy statistics. Technical report, ESDS International, University of Manchester, 2011b.
- M. Z. Jacobson and M. A. Delucchi. Providing all global energy with wind, water, and solar power, part i: Technologies, energy resources, quantities and areas of infrastructure, and materials. *Energy Policy*, 39(3):1154 – 1169, 2011. ISSN 0301-4215. doi: 10.1016/j.enpol.2010.11.040.
- W. Ji and K. C. Chee. Prediction of hourly solar radiation using a novel hybrid model of arma and tdnn. *Solar Energy*, 85(5):808–817, 2011. ISSN 0038092x. doi: 10.1016/j.solener.2011.01.013.

- T. Jónsson, P. Pinson, H. A. Nielsen, H. Madsen, and T. Nielsen. Forecasting electricity spot prices accounting for wind power predictions. *IEEE Transactions on Sustainable Energy*, Submitted:, 2012.
- R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.
- R. Koenker. *quantreg: Quantile Regression*, 2011. URL <http://CRAN.R-project.org/package=quantreg>. R package version 4.76.
- N. R. Kristensen, H. Madsen, and S. B. Jørgensen. Parameter estimation in stochastic grey-box models. *Automatica*, 40(2):225 – 237, 2004. ISSN 0005-1098. doi: DOI:10.1016/j.automatica.2003.10.001.
- D. Lew, N. Miller, K. Clark, G. Jordan, and Z. Gao. Impact of high solar penetration in the western interconnection. *Contract*, 303:275–3000, 2010.
- E. Lorenz, D. Heinemann, H. Wickramaratne, H. Beyer, and S. Bofinger. Forecast of ensemble power production by grid-connected pv systems. In *Proc. 20th European PV Conference, September 3-7, 2007, Milano*, 2007.
- E. Lorenz, T. Scheidsteger, J. Hurka, D. Heinemann, and C. Kurz. Regional pv power prediction for improved grid integration. *Progress in Photovoltaics: Research and Applications*, 19(7):757–771, 2011. ISSN 1099-159X.
- H. Madsen. *Time Series Analysis*. Chapman & Hall, 2007.
- H. Madsen and J. Holst. Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings*, 22(1):67–79, 1995. ISSN 03787788.
- B. V. Mathiesen, H. Lund, and K. Karlsson. 100growth. *Applied Energy*, 88(2): 488 – 501, 2011. ISSN 0306-2619. doi: 10.1016/j.apenergy.2010.03.001.
- P. Meibom, J. Kiviluoma, R. Barth, H. Brand, C. Weber, and H. V. Larsen. Value of electric heat boilers and heat pumps for wind power integration. *Wind Energy*, 10(4):321–337, 2007. ISSN 1099-1824. doi: 10.1002/we.224.
- M. Nicolosi and C. Nabe. The long-term effects of high shares of pv in the power system—an analysis of the german power market. In *1st International Workshop on the Integration of Solar Power into Power Systems*, 2011.
- H. Nielsen, T. Nielsen, A. Joensen, H. Madsen, and J. Holst. Tracking time-varying-coefficient functions. *International Journal of Adaptive Control and Signal Processing*, 14(8):813–828, 2000.
- H. A. Nielsen, H. Madsen, and T. S. Nielsen. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. *Wind Energy*, 9(1-2):95–108, 2006. ISSN 1099-1824. doi: 10.1002/we.180.

- P. Palensky and D. Dietrich. Demand side management: Demand response, intelligent energy systems, and smart loads. *Industrial Informatics, IEEE Transactions on*, 7(3):381–388, aug. 2011. ISSN 1551-3203. doi: 10.1109/TII.2011.2158841.
- T. Pedersen, P. Andersen, K. Nielsen, H. Starmose, and P. Pedersen. Using heat pump energy storages in the power grid. In *Control Applications (CCA), 2011 IEEE International Conference on*, pages 1106–1111, sept. 2011. doi: 10.1109/CCA.2011.6044504.
- S. Pelland, G. Galanis, and G. Kallos. Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model. *Progress in Photovoltaics: Research and Applications*, pages n/a–n/a, 2011. ISSN 1099-159X. doi: 10.1002/pip.1180.
- B. Perers. An improved dynamic solar collector test method for determination of non-linear optical and thermal characteristics with multiple regression. *Solar Energy*, 59(4-6):163–178, 1997. ISSN 0038092x.
- B. Perers, S. Furbo, J. Fan, E. Andersen, and Z. Chen. Solar combisystems with forecast control to increase the solar fraction and lower the auxiliary energy cost. In *ISES Solar World Congress 2011 Proceedings*, page , 2011. ISBN 978-3-9814659-0-7. Presented at: ISES Solar World Congress, SWC ; 30 : Kassel, Germany, 2011.
- P. Pinson, C. Chevallier, and G. Kariniotakis. Trading wind generation from short-term probabilistic forecasts of wind power. *Power Systems, IEEE Transactions on*, 22(3):1148–1156, aug. 2007a. ISSN 0885-8950. doi: 10.1109/TPWRS.2007.901117.
- P. Pinson, H. A. Nielsen, J. K. Møller, H. Madsen, and G. N. Kariniotakis. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy*, 10(6):497–516, 2007b. ISSN 1099-1824. doi: 10.1002/we.230. URL <http://dx.doi.org/10.1002/we.230>.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL <http://www.R-project.org/>. ISBN 3-900051-07-0.
- C. Rigollier, O. Bauer, and L. Wald. On the clear sky model of the esra - european solar radiation atlas - with respect to the heliosat method. *Solar Energy*, 68(1):33–48, 2000. ISSN 0038-092X. doi: 10.1016/S0038-092X(99)00055-9.
- S. Sayeef, T. Moore, S. Percy, D. Cornforth, J. Ward, and D. Rowe. Characterisation and integration of high penetration solar power in australia – a solar intermittency study. In *1st International Workshop on the Integration of Solar Power into Power Systems Aarhus, Denmark, 24 October 2011*, 2011.

- J. Schmelter, M. Lange, and U. Focken. Weather class depending combinations of solar power forecasts – operational experiences. In *1st International Workshop on the Integration of Solar Power into Power Systems Aarhus, Denmark, 24 October 2011*, 2011.
- G. Y. Yang, J. Østergaard, S. B. Kjær, A. Constantin, H. P. Ballegaard, R. D. Lazar, U. Borup, C. Stephansen, M. Mattesen, and J. B. Sørensen. Smart integration of photovoltaic power systems on the island of bornholm. In *1st International Workshop on the Integration of Solar Power into Power Systems Aarhus, Denmark, 24 October 2011*, 2011.
- S. Younes, R. Claywell, and T. Muneer. Quality control of solar radiation data: Present status and proposed new approaches. *Energy*, 30(9):1533 – 1549, 2005. ISSN 0360-5442. doi: 10.1016/j.energy.2004.04.031. Measurement and Modelling of Solar Radiation and Daylight- Challenges for the 21st Century.
- Y. Zong, D. Kullmann, A. Thavlov, O. Gehrke, and H. Bindner. Active load management in an intelligent building using model predictive control strategy. In *PowerTech, 2011 IEEE Trondheim*, pages 1 –6, june 2011. doi: 10.1109/PTC.2011.6019347.
- K. Zweibel, J. Mason, and V. Fthenakis. A solar grand plan. *Scientific American*, 298(1):64–73, 2008.
- J. Široký, F. Oldewurtel, J. Cigler, and S. Prívara. Experimental analysis of model predictive control for an energy efficient building heating system. *Applied Energy*, 88(9):3079 – 3087, 2011. ISSN 0306-2619. doi: 10.1016/j.apenergy.2011.03.009.

P A P E R A

A non-parametric method for correction of global radiation observations

Authors:

Peder Bacher¹, Henrik Madsen¹, Bengt Perers², and Henrik Aalborg Nielsen³

Submitted to:

Solar Energy (April 2012)

¹DTU Informatics, Richard Pedersens Plads, Building 321, DK-2800 Lyngby, Denmark

²DTU Civil Engineering, Technical University of Denmark, DK-2800, Lyngby, Denmark

³ENFOR A/S, Lyngsø Allé 3, DK-2970 Hørsholm, Denmark (URL: www.enfor.eu)

Abstract

This paper presents a method for correction and alignment of global radiation observations based on information obtained from calculated global radiation, in the present study one-hour forecast of global radiation from a numerical weather prediction (NWP) model is used. Systematical errors detected in the observations are corrected. These are errors such as: tilt in the leveling of the sensor, shadowing from surrounding objects, clipping and saturation in the signal processing, and errors from dirt and wear. The method is based on a statistical non-parametric clear-sky model which is applied to both the observed and the calculated radiation in order to find systematic deviations between them. The method is applied to correct global radiation observations from a climate station located at a district heating plant in Denmark. The results are compared to observations recorded at the Danish Technical University. The method can be useful for optimized use of solar radiation observations for forecasting, monitoring, and modeling of energy production and load which are affected by solar radiation.

Nomenclature

G_t	Observed global radiation [W/m ²]
G_t^{nwp}	Numerical weather predictions (NWP) of global radiation [W/m ²]
G_{cs}	Clear-sky global radiation [W/m ²]
B_{cs}	Direct clear-sky global radiation [W/m ²]
D_{cs}	Diffuse clear-sky global radiation [W/m ²]
G	Global radiation [W/m ²]
I_{ext}	Extraterrestrial radiation [W/m ²]
G_t^{pr}	Projection of global radiation to the plane normal to the direct solar radiation [W/m ²]
$\hat{G}_t^{\text{pr,cs}}$	Estimated clear-sky radiation on a plane normal to the direct solar radiation [W/m ²]
\hat{G}_t^{cs}	Estimated clear-sky global radiation (modeled based on observations) [W/m ²]
$\hat{G}_t^{\text{nwp,cs}}$	Clear-sky global radiation for numerical weather predictions (NWP) [W/m ²]
\hat{G}_t^{co}	Corrected global radiation [W/m ²]
θ_t^{zenith}	Solar zenith angle [rad]
$\tau_{\text{a,B}}$	Transmittance function of the atmosphere for direct radiation under clear-sky conditions
τ_{c}	Transmittance function of clouds in the atmosphere
β_t	Parameter vector for the local quantile regression
$\rho_q(u)$	The quantile regression objective function
q	Sample quantile to be estimated in the local quantile regression
i	Counter of days [days]
j	Counter in samples
t	Time [hours]
t_{sp}	Sample period [hours]
h_{doy}	Bandwidth of kernel function in the <i>day of year</i> dimension [days]
h_{tod}	Bandwidth of kernel function in the <i>time of day</i> dimension [hours]

A.1 Introduction

The transition to a reliable and secure energy system based on weather dependent production technologies, especially wind and solar, will require new methods for automated handling of climate data recorded at, in most cases, unsupervised and uncalibrated stations. Reliable observations of solar radiation

are an important source of information for operation of the energy system, especially for the energy production and load which are dependent on the solar radiation, for example production from photovoltaics and solar collectors, and load from heating and cooling of buildings.

Observations of solar radiation are exposed to many sources of errors. Younes et al. (2005) list the most important types of errors and divide the errors into two major categories: equipment errors and operation related errors. The present solar radiation sensor technology makes it easy and cheap to install and connect sensors to the Internet, both for professional and amateur applications. Web sites already provide on-line data (DMI, 2012), which can become an important source of information for operation of energy systems. Such, mostly unsupervised and unvalidated installations, will be highly exposed to different error sources.

In the present study observations of global radiation from a station at a district heating plant in Sønderborg, Denmark, are used. Three types of errors are found in the observations: tilt in the leveling of the sensor, shadowing from surrounding objects, and clipping at a maximum level. A method is presented for correction of the observations on the basis of information extracted from global radiation calculated using a model based on physical principles. The method is based on a non-parametric statistical clear-sky model and requires no further information about the installation and sensor than the observed values and the location of the station. With the statistical clear-sky model the sensor output level under clear-sky conditions is modeled directly from the observations. This is compared to solar radiation calculated with a clear-sky model based on physical modeling of the optical effects through the atmosphere, such as the models described by Davies and McKay (1982), Bird (1984), Rigollier et al. (2000), Mueller et al. (2004), and Ineichen (2006). In the present study forecasts from a numerical weather prediction (NWP) model is used. The result after correction of the observations is compared to high quality measurements recorded at the Danish Technical University.

Studies on quality control of measured solar radiation data can be found in the literature. The procedures are semi-automatic and are mostly based on comparison to physical models for detection of erroneous measurements (Geiger et al., 2002), (Younes et al., 2005), (Isaac and Moradi, 2009) and (Journée and Bertrand, 2011).

The paper is organized as follows: the data used in the study is presented in the next section. This is followed by a section in which the statistical clear-sky model is described and a section where the correction is presented. The paper ends with a discussion of the method and a conclusion.



Figure A.1: The weather station in Sønderborg, which is mounted on a pole on the roof of a single-storey district heating plant building (in the image it is on the left side of the building).

A.2 Data: Observations and numerical weather predictions of global radiation

The data used in this study consists of time series of global radiation observed at two weather stations: one located in Sønderborg (54.91°N and 9.80°E) and one located at DTU Byg in Lyngby (55.79°N and 12.52°E), both in Denmark. In addition NWP's of global radiation for the same locations are used. All values are hourly averages. All times are in UTC and the time points are set to the end of the hour.

A.2.1 Observations

The observations from Sønderborg are recorded with a weather station, which is located at a district heating plant. The weather station is mounted on a pole on a single-storey building as seen on the image in Figure A.1. No information about the type of the solar radiation sensor was available. The time series from Sønderborg is

$$\{G_t; t = 1, \dots, N\} \quad (\text{A.1})$$

where $N = 17520$ and G_t is the observed average global radiation between time t and $t - 1$. The upper plot in Figure A.2 shows the series which spans from

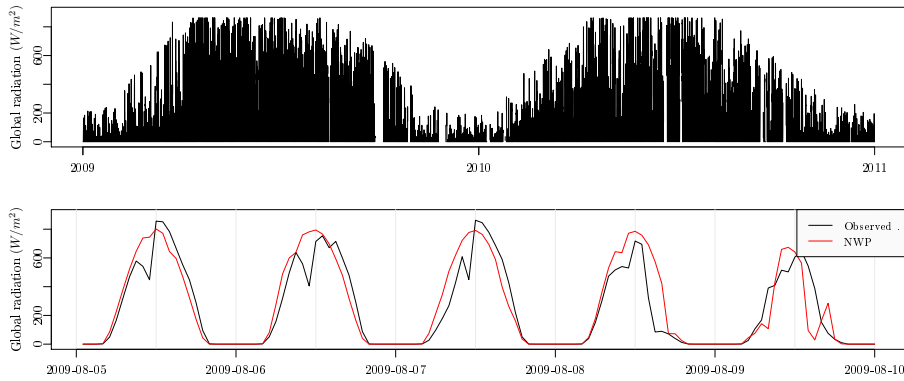


Figure A.2: The upper plot shows the time series of observed global radiation in Sønderborg. In the lower plot the observations and NWP's of global radiation in Sønderborg are shown for five days in August 2009.

2009-01-01 to 2011-01-01. From this plot it is readily seen that the observations are not without systematic errors, for example it can be seen that the values are clipped at a maximum level. This and other types of systematic errors are corrected for the Sønderborg observations using the method described in this paper.

The second series of observed global radiation is from a weather station at DTU Byg in Lyngby and is used as a reference to check the corrected data. The upper plot in Figure A.3 shows the series which spans from 2009-01-01 to 2010-01-01. It was measured with a Kipp & Zohnen CM10 pyranometer and the weather station was regularly supervised in the measuring period. The measurement error is in the range of maximum $\pm 3\%$ from the world standard and high class calibrated sensor inter-comparisons indicate an error within the range of $\pm 1\%$. The lower plot in Figure A.3 shows the observations together with the NWP's of global radiation (defined in the next section) for five days in August 2009. It is seen that the level of the observed global radiation is generally lower than the level of the NWP's, but that this there is no systematic difference between the deviation in the morning and in the afternoon. The lower level is most likely due to a bias in the NWP's. Since the accuracy of the DTU observations is high and no systematic errors, apart from the generally lower level, is seen, then it is found valid to assume that the DTU observations can be used as a reference to verify the NWP's and the results of the correction.

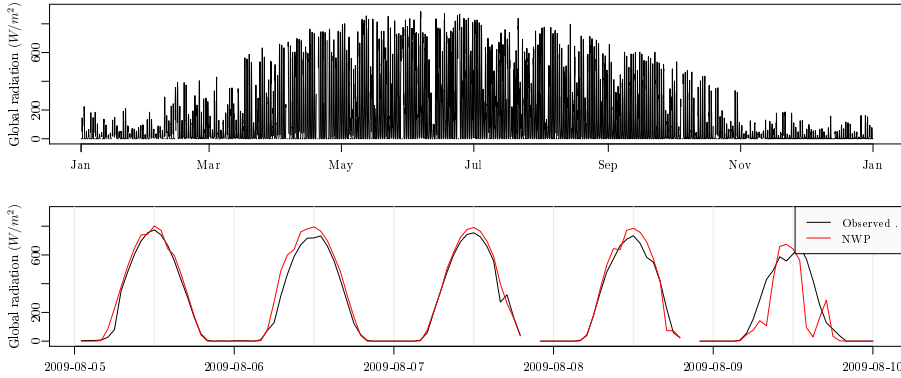


Figure A.3: The upper plot shows the time series of observed global radiation at DTU Byg covering the entire year 2009. In the lower plot the observations and NWP of global radiation at DTU Byg are shown for five days in August 2009.

A.2.2 Numerical weather predictions

The numerical weather predictions (NWP) used in the study are provided by the Danish Meteorological Institute (DMI). The NWP model used is DMI-HIRLAM-S05, which has a 5 kilometer grid and 40 vertical layers, see (DMI, 2011) and (Hansen Sass et al., 2002) for more details. The forecasts are updated four times per day and have a calculation delay of 4 hours (e.g. the forecast starting at 00:00 is available at 04:00). Two time series, consisting of the latest available forecast (lead times are 5 to 11 hours) of global radiation, are used: one for the location in Sønderborg and one for the location of DTU in Lyngby. The time series of NWP for the Sønderborg location is used for the correction. It is denoted with

$$\{G_t^{\text{nwp}}; t = 1, \dots, N\} \quad (\text{A.2})$$

The time series for DTU Byg in Lyngby is shown, together with the observations, in the lower plot of Figure A.3 for five days in August.

A.2.3 Systematic errors in Sønderborg observations

The lower plot in Figure A.2 shows the Sønderborg observations and the NWP of global radiation for five days in August 2009. From the first day, which is a clear-sky day, at least two types of errors can be seen in the observations:

compared to the NWP the observed level is too low in the morning and too high in the afternoon, which is most likely due to the sensor being tilted. It could also be due to a shift in time of the sensor, however it was thoroughly checked that the night hours, where the radiation was zero (or very close to zero), are with only a few exceptions the same hours for both the observed and the NWP, indicating that they are well synchronized. The second type of error is seen just before noon, where the observations have a drop, which is repeated at the same *time of day* on following clear-sky day. The drop is caused by shading from the chimney, which is located close to the weather station, as seen on the image in Figure A.1.

The scatter plot in Figure A.4 shows the observed values versus the NWP, together with two lines indicating the relation between the variables in the morning and in the afternoon. The lines are calculated using locally weighted least squares regression between the observations and the NWP, using the function `loess()` in R (R Development Core Team, 2011) with a bandwidth: `span=0.9`. A similar plot for the DTU observations is found in Figure A.5. The following three distinct systematic errors can be seen from the scatter plot for the Sønderborg observations:

1. Firstly, the observations are clipped at a maximum level around 860 W/m².
2. Secondly, the level of the morning observations is generally lower than the level of the afternoon observations. This is confirmed by the fitted regression lines, which mostly have a difference of at least 50 to 75 W/m². This is clearly a larger difference than seen for the two fitted lines for the DTU observations in Figure A.5.
3. Finally, the morning values are significantly lower in the NWP range of 700 to 900 W/m². These values are the observations in the drop before noon, which, as described earlier, is caused by shadowing from the chimney right next to the weather station.

Considering the scatter plot for the DTU observations in Figure A.5 it is seen that these systematic errors are not found in the DTU observations. As noted before the level of the DTU observations is generally a bit lower than the level of NWP, which is most likely due to a bias of the NWP, since the accuracy of the DTU observations is verified to be in the range of $\pm 3\%$. For correction of the systematic errors, as the listed above, a statistical clear-sky model fitted to the observations can be used, as outlined in the following sections.

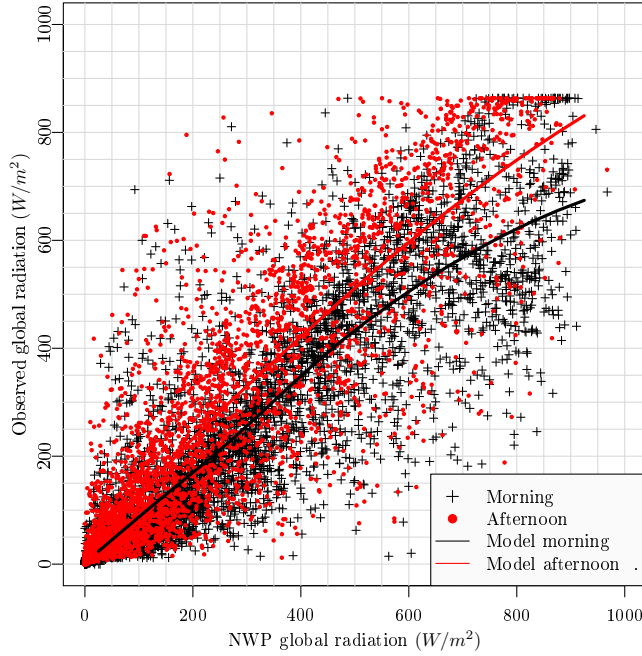


Figure A.4: The values of the Sønderborg observations versus the NWP's covering the entire year 2009. The morning values and the afternoon values are indicated by different symbols and colors. The two lines show a locally weighted least squares regression estimate of the relation between the variables in the morning and the afternoon.

A.3 Statistical clear-sky model

In this section it is described how the clear-sky global radiation is modeled using a statistical model. With the statistical clear-sky model the level under clear-sky conditions at time t is estimated for the particular series of observations. It is the output of the sensor under clear-sky conditions which is estimated. This implies that if an observation is affected by a systematical error, for example shadowing from an object in the surroundings, the estimated clear-sky output will be lowered. It is this feature which enables the model to be used for correction. The statistical clear-sky model is a non-parametric model based on local polynomial quantile regression (Koenker, 2005) similar to the clear-sky model presented in (Bacher et al., 2009).

Usually, clear-sky models are models with which the global radiation in clear

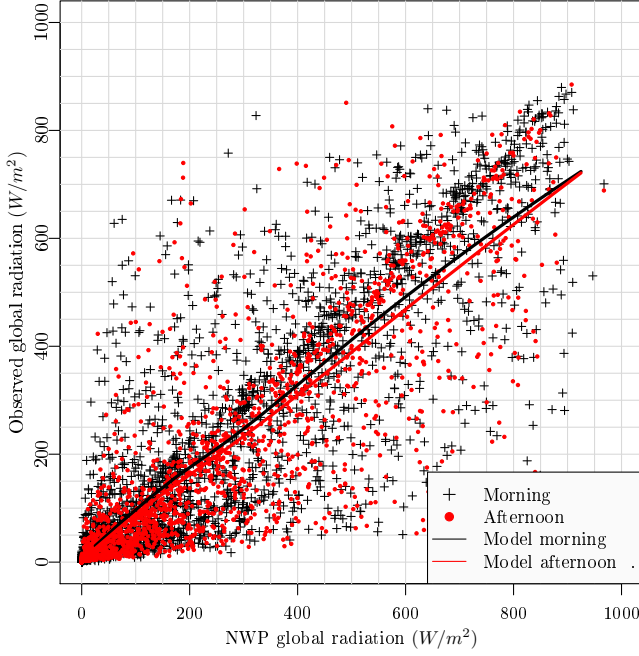


Figure A.5: The values of observed versus NWP global radiation at DTU Byg in Lyngby, Denmark. The morning values and the afternoon values are plotted with different symbols and colors. The two lines are indicating the relation between the variables: one for the morning and one for the afternoon.

(non-overcast) sky at any given time can be calculated based on physical modeling of the atmosphere. Usually the clear-sky global radiation G_{cs} is separated into a direct (or beam) B_{cs} and diffuse D_{cs} component

$$G_{cs} = B_{cs} + D_{cs} \quad (\text{A.3})$$

which are then modeled separately. The direct component by

$$B_{cs} = I_{\text{ext}} \cos(\theta_{\text{zenith}}) \tau_{a,B} \quad (\text{A.4})$$

where I_{ext} is the extraterrestrial radiation, θ_{zenith} is the solar zenith angle and $\tau_{a,B}$ is a transmittance function of the atmosphere for direct radiation under clear-sky conditions, which for example can be modeled taking Rayleigh scattering, aerosol extinction, and ozone, water and uniformly mixed gas absorption into account Bird and Riordan (1984).

The diffuse component can be modeled by adding several contributions from

reflections and scattering through the atmosphere.

The global radiation (at the surface of the earth) can be modeled by

$$G = G_{cs} \tau_c \quad (\text{A.5})$$

where τ_c is a transmittance function of clouds in the atmosphere, which can be modeled with "layer models" (Davies and McKay, 1982) where cloud layer transmittance and reflections are taken into account.

A clear-sky model, similar to the one proposed by Bacher et al. (2009) for observations of solar power, is here proposed for observations of global radiation. The proposed clear-sky model does not include any prior physical knowledge, it is based solely on the information obtained from the observations. It is denoted as a statistical clear-sky model, since it is based on a non-parametric statistical model of clear-sky radiation. Information embedded in the observations, which is particular for the sensor and its location, can be modeled with the statistical clear-sky model, for example shadowing and non-horizontal leveling of the sensor. This is a fundamental difference to the clear-sky models based on prior physical knowledge, which implies that the statistical clear-sky model can be used for different applications.

The statistical clear-sky model is based on time series of global radiation observations (or simulated values) and is defined by

$$G_t = \hat{G}_t^{cs} \tau_t \quad (\text{A.6})$$

where the t is used to indicate that the variables the time series of actual observations, G_t is observed global radiation, \hat{G}_t^{cs} is estimated clear-sky global radiation and τ_t is a factor, which is much to alike τ_c , but different due to the fact that it is estimated based on information from observations and not calculated based on prior physical knowledge. It is noted here that the clear-sky model could be defined for the direct component solely, which would be obvious since nearly all local systematic effects have a much higher impact on the direct component compared to the diffuse component. However, since the application of the clear-sky model in the present study is for observations of global radiation and since the systematic errors would propagate into both the direct and diffuse component calculated with a splitting scheme, such as suggested by Ruiz-Arias et al. (2010), the clear-sky model is applied to the global radiation directly.

Considering the observed global radiation as samples of a random variable with a probability distribution function, which is a function of the *day of year* x_t and the *time of day* y_t , the observed clear-sky global radiation can be estimated as a quantile

$$\hat{G}_t^{cs} = Q_q(x_t, y_t) \quad (\text{A.7})$$

54 A non-parametric method for correction of global radiation observations

of this distribution function, where the quantile $q \in [0, \dots, 1]$ must be close to one

$$q \lesssim 1 \quad (\text{A.8})$$

Assuming that the quantile function is a smooth function it can be approximated with local quantile regression Koenker (2005). The result in the three-dimensional space formed by global radiation, *day of year* and *time of day*, can be seen as a surface which follows the observed global radiation under clear-sky conditions and is located "on top" of the point cloud of observed global radiation.

In order to decrease the gradient and curvature of the estimated clear-sky radiation surface a projection is carried out. The projection is from the horizontal plane to the plane which is normal to the direct solar radiation (i.e. the plane tracking the sun position)

$$G_t^{\text{pr}} = \frac{G_t}{\cos(\theta_t^{\text{zenith}})} \quad (\text{A.9})$$

where θ_t^{zenith} is the average solar zenith angle in the sample period between $t - 1$ and t . Values where $\cos(\theta_t^{\text{zenith}}) < 0.01$ are removed: this corresponds to sun elevation below 0.5° . The quantile close to one is then estimated for the projected values. A general form of the proposed statistical clear-sky model is formulated in A.7, which is based on a local quantile regression model with second order polynomials and a two-dimensional kernel in both the *day of year* and *time of day* dimensions.

For correction of hourly values a local quantile regression model based only on a one-dimensional kernel, where on the *day of year* dimension is used, was found most suitable. The reason for using only a one-dimensional kernel, and not including the *time of day* dimension in the local weighting, is that the model becomes too biased and the estimated clear-sky global radiation does not follow the drop before noon caused by shadowing (the systematic error described on page 50) very well. Hence only values lagged in steps of 24 hours from t are used as input, which is a similar approach as in classical decomposition of seasonal time series (Cleveland and Tiao, 1976). Furthermore, it is noted that this is equivalent to using a bandwidth in the *time of day* dimension below one hour (i.e. below the sample period) for the two-dimensional model presented in A.7, hence for time series with a shorter sample period a two-dimensional model should be considered. The applied local quantile regression model based on a third order polynomial is

$$\hat{\beta}_t = \underset{\beta \in \mathbb{R}^4}{\operatorname{argmin}} \sum_{i=-\infty}^{\infty} \rho_q(G_{t+24i}^{\text{tr}} - (\beta_{0,t} + \beta_{1,t}i + \beta_{2,t}i^2 + \beta_{3,t}i^3))K(i) \quad (\text{A.10})$$

where $\rho_q(u) = u(q - I(u < 0))$ is the quantile regression objective function (see (Koenker, 2005) and (Koenker, 2011)), $q \in [0, \dots, 1]$ is the sample quantile to be estimated, $i \in \mathbb{N}$ is a counter of days, and $K(i)$ is a kernel function. The estimated projected clear-sky radiation is then found as the local intercept

$$\hat{G}_t^{\text{pr,cs}} = \hat{\beta}_{0,t} \quad (\text{A.11})$$

The weights are calculated with the Epanechnikov kernel function

$$K(i) = \begin{cases} \frac{3}{4} \left(1 - \left[\frac{|i|}{h_{\text{day}}} \right]^2 \right) & \text{for } \frac{|i|}{h_{\text{day}}} \geq 1 \\ 0 & \text{for } \frac{|i|}{h_{\text{day}}} < 1 \end{cases} \quad (\text{A.12})$$

where h_{day} is the bandwidth.

The R package **quantreg** implementation of quantile regression was used for the estimation (Koenker, 2011). Finally, the estimated projected clear-sky radiation on the projected plane is projected back to the horizontal plane by

$$\hat{G}_t^{\text{cs}} = \hat{G}_t^{\text{pr,cs}} \cos(\theta_t^{\text{zenith}}) \quad (\text{A.13})$$

Finally, in order to take the clipping at a maximum level into account, the estimated clear-sky radiation is limited to the maximum value of the observations

$$\hat{G}_t^{\text{cs}} = \begin{cases} \hat{G}_t^{\text{cs}} & \text{for } \hat{G}_t^{\text{cs}} \leq G_t^{\text{max}} \\ G_t^{\text{max}} & \text{for } \hat{G}_t^{\text{cs}} > G_t^{\text{max}} \end{cases} \quad (\text{A.14})$$

where G_t^{max} is the maximum value of global radiation observations.

The selection of suitable values for the parameters (here the quantile and the kernel bandwidth) for the fitting of the local quantile regression model, would preferably be based on a measure of performance for estimation clear-sky global radiation. Then the parameters could be optimized in order to achieve the best performance. However thorough studies are required in order to define such a measure. Therefore the parameter values are selected based on visual inspection of the estimated clear-sky global radiation for days with only clear-sky. These days are chosen such that they are distributed evenly over the entire period. The selected values are

$$q = 0.97, \quad h_{\text{day}} = 125 \quad (\text{A.15})$$

which gives the estimate of the clear-sky global radiation for the observations \hat{G}_t^{cs} shown in Figure A.6 and for the NWP's $\hat{G}_t^{\text{nwp,cs}}$ shown in Figure A.7. Note, that the estimated surface for the observations is clipped at the maximum value of the observations, which gives the "flat" top. Furthermore, notice that the drop due to shadowing is clearly seen in the estimated clear-sky radiation for the observations.

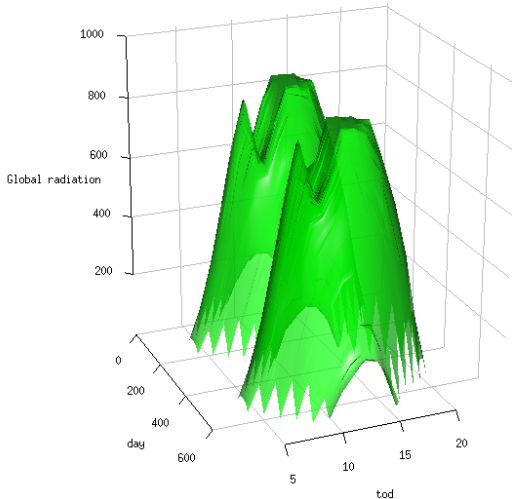


Figure A.6: The clear-sky global radiation estimated for the S nderborg observations. Shown as a surface parametrized in the two dimensions: day (days since 2009-01-01) and tod (*time of day*).

A.4 Correction of observations

The correction of the observations is carried out by multiplying the observations with the ratio between the estimated clear-sky radiation for the NWP and the observations

$$\hat{G}_t^{\text{co}} = \frac{\hat{G}_t^{\text{nwp,cs}}}{\hat{G}_t^{\text{cs}}} G_t \quad (\text{A.16})$$

The level of the correction applied, i.e. $\hat{G}_t^{\text{nwp,cs}}/\hat{G}_t^{\text{cs}}$, is shown as function of days since 2009-01-01 and the *time of day* in Figure A.8. The systematical error caused by a tilt of the sensor, resulting in a too low level of the observations in the morning and too high level in the afternoon, can be directly seen in the correction, since in the morning the correction is generally above one and the afternoon level below one. Also apparent is the drop in the observed level due to shadowing objects, especially seen between 9 to 10 am.

The corrected observations are plotted versus the NWPs in Figure A.9, including the local least squares estimate of the relation in the morning and in the afternoon. This plot is similar to the plot in Figure A.4. By comparison of

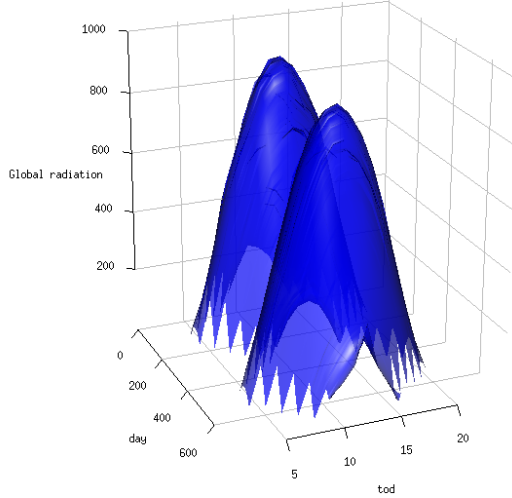


Figure A.7: The clear-sky global radiation estimated for the NWP for Sønderborg. Shown as a surface parametrized in the two dimensions: day (days since 2009-01-01) and tod (*time of day*).

the two plots it is seen that the difference between the estimated relation in the morning and the estimated relation in the afternoon has been decreased significantly. A visual comparison to the similar plot of the high quality DTU observations in Figure A.5 verifies that the pattern of the scatter after the correction is much closer to the pattern found there. It can also be seen that the clipping at a maximum level has been corrected. Finally, it is found that the overall scattering has been reduced. This is confirmed by a comparison of the errors for an estimated relation similar to the ones in Figure A.4 and A.9, but using all data points (except nighttime values), i.e. no distinction between morning and afternoon. Note here that this measure is only used to give a rough indication of the performance of the correction. The root mean square error (RMSE) and mean absolute error (MAE) before the correction are

$$\text{RMSE}_{\text{before}} = 114 \text{ W/m}^2, \quad \text{MAE}_{\text{before}} = 79 \text{ W/m}^2 \quad (\text{A.17})$$

and after the correction

$$\text{RMSE}_{\text{after}} = 101 \text{ W/m}^2, \quad \text{MAE}_{\text{after}} = 67 \text{ W/m}^2 \quad (\text{A.18})$$

Hence a notably reduction in RMSE and MAE is achieved by applying the correction.

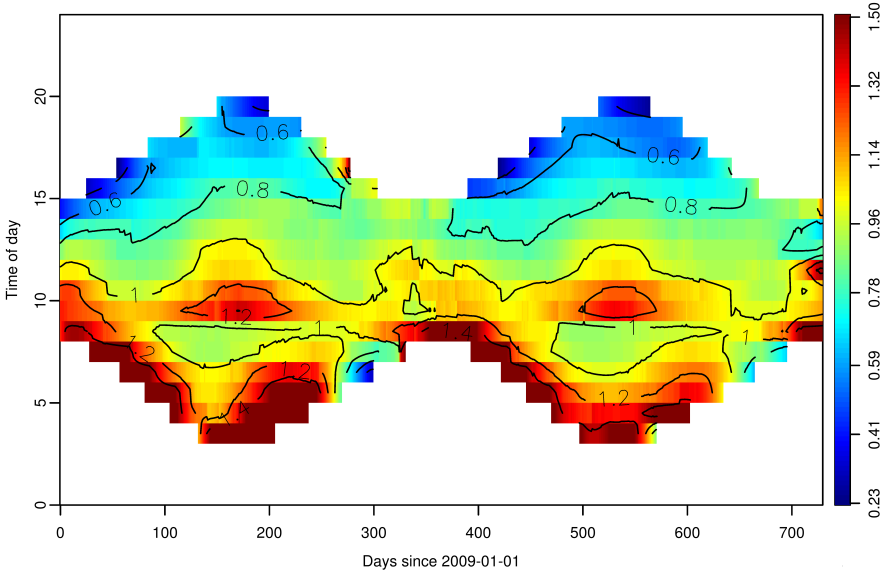


Figure A.8: The applied correction, which is the ratio between the estimated clear-sky radiation of the observations and the NWP, as a function of days and *time of day*.

A.4.1 On-line operation

For on-line operation the model has to be applied causally, such that only past values can be used for the correction. A causal correction was calculated with slightly different parameter values for the clear-sky model, again selecting the parameters from visual inspection. The estimated quantile q was decreased and the kernel bandwidth h_{doy} increased slightly to

$$q = 0.95, \quad h_{\text{doy}} = 150 \quad (\text{A.19})$$

Using a one-sided kernel will increase the bias of the estimates, which is also reflected by a slightly increased RMSE and MAE of the `loess` fit for the corrected observations to

$$\text{RMSE}_{\text{after}} = 103 \text{ W/m}^2, \quad \text{MAE}_{\text{after}} = 67 \text{ W/m}^2 \quad (\text{A.20})$$

Considering the similar plots as presented for the causal correction showed only a small visual difference. Hence it is found that the method works well for on-line operation.

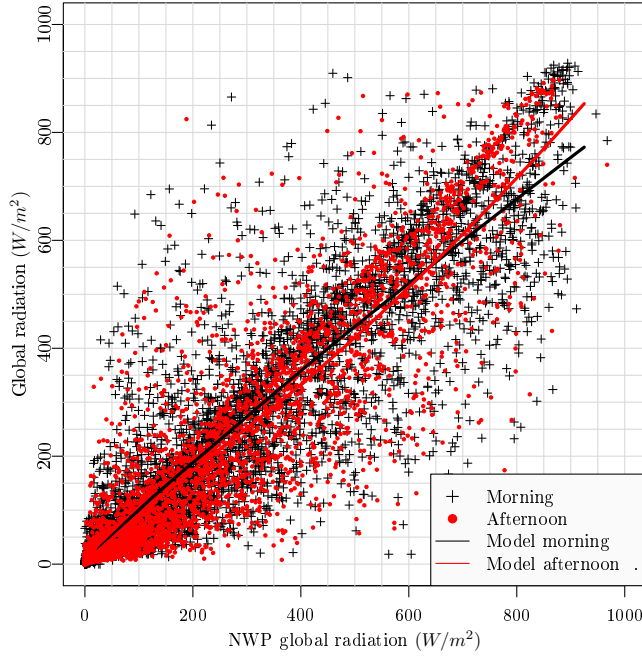


Figure A.9: The values of the corrected observations versus NWP's of global radiation. The morning values and the afternoon values are indicated by different symbols and colors. The two lines show a locally weighted least squares regression estimate of the relation between the variables in the morning and the afternoon.

A.5 Discussion

In this section the correction method and results are discussed together with considerations on how to improve the method.

Considering the fitting of the clear-sky model it is noted that the model which should be applied, is dependent on the time resolution of the data. For resolutions higher than hourly a two-dimensional kernel, which also includes the *time of day* dimension, should be used, as in (Bacher et al., 2009) where a similar clear-sky model was applied to fifteen minutes values. Regarding the parameters needed to be tuned in the clear-sky model - the quantile, kernel bandwidth, and order of the polynomial - some manual interaction is required. However the method could be based on a parameter optimization criteria, hence an objective measure to evaluate the performance of the correction, possibly based

on cross validation (Friedman et al., 2001), and applied automatically for the general case. Hence it can also be used for monitoring and data quality classification for sub-daily solar data. Clearly, an objective measure of performance of the correction is needed in order to further develop and improve the correction method.

Improvements of the method could be formed by combining it with a prior step in which a parametric model is fitted to correct for drift in time and tilt in the leveling of the sensor. Another possibility for improvement is to treat the direct and diffuse radiation separately, since most of the systematic errors, for example tilt and shadowing, will have a different impact on direct and diffuse radiation. This will require, if the direct and diffuse are not measured separately, a splitting into a diffuse and direct component, which could be carried out with a scheme such as suggested by Ruiz-Arias et al. (2010) and Duffie and Beckman (2006, p. 75-77). However applying such a scheme will cause the effect of the systematic errors to propagate into both the direct and the diffuse components. Another approach would be to enhance the correction method by using more than one quantile, in the presented approach only a single quantile close to one is used. Several quantiles can be estimated for both the observed and calculated radiation, which, together with an interpolation scheme, will form a more extensive correction. Clearly this also requires that the calculated global radiation, i.e. here the NWP, describes the distribution well over the entire range of global radiation. Finally, it is mentioned that for on-line operation the method can be implemented computationally very efficient using time-adaptive quantile regression (Møller et al., 2008).

A.6 Conclusion

A correction method based on statistical non-parametric modeling techniques is presented and applied on hourly observations of global radiation. Several typical errors in the observations can be corrected with the method, including: tilt in the leveling of the sensor, shadowing from objects in the surroundings and clipping of the observations at a maximum level. The method works semi-automatically and no prior information about the sensor and its surroundings, besides the observations and location, is required. Furthermore only a few parameters need to be tuned. Information embedded in NWP of global radiation is used for the correction, but this could be replaced with any calculated clear-sky global radiation model. The method is well suited as part of monitoring and operation applications for which local solar radiation observations provide valuable information, e.g. for forecasting of climate dependent renewables such as solar thermal, PV and heating systems. Finally, it is briefly discussed how

the method can be improved or extended in several ways.

Acknowledgement

The observations of global radiation was very kindly provided by Sønderborg Fjernvarme and we also thank the Danish Meteorological Institute for making their numerical weather predictions available. Finally, acknowledgments are given to the The Danish Council for Strategic Research, which have provided the financial support for the project "Solar/electric heating systems in the future energy system" under which the work was carried out.

A.7 Two-dimensional local statistical clear-sky model

The proposed statistical clear-sky model in a general form is described in this section. It is based on a two-dimensional second-order polynomial local quantile regression model. In this form the local weighting is carried out with a two-dimensional multiplicative kernel function in the *day of year* and *time of day* dimensions.

The model

$$\hat{\beta}_t = \underset{\beta \in \mathbb{R}^5}{\operatorname{argmin}} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \rho_q(G_{t+\frac{24}{t_{\text{sp}}}}^{\text{tr}}{}_{i+j} - (\beta_{0,t} + \beta_{1,t}i + \beta_{2,t}i^2 + \beta_{3,t}j + \beta_{4,t}j^2))K(i,j) \quad (\text{A.21})$$

where t_{sp} is the sample period of the time series in hours, $\rho_q(u) = u(q - I(u < 0))$ is the quantile regression objective function (see (Koenker, 2005) and (Koenker, 2011)), $q \in [0, \dots, 1]$ is the sample quantile to be estimated, $i \in \mathbb{N}$ is a counter of days, $j \in \mathbb{N}$ is a counter in steps of the sample period, and $K(i, j)$ is a kernel function. The model could easily be reduced or expanded to polynomials of different orders. The estimated projected clear-sky radiation is then found as the local intercept

$$\hat{G}_t^{\text{pr,cs}} = \hat{\beta}_{0,t} \quad (\text{A.22})$$

The weights are calculated with the Epanechnikov kernel function

$$K(i, j) = \begin{cases} \frac{9}{16} \left(1 - \left[\frac{|i|}{h_{\text{doy}}}\right]^2\right) \left(1 - \left[\frac{|j|t_{\text{sp}}}{h_{\text{tod}}}\right]^2\right) & \text{for } \frac{|i|}{h_{\text{doy}}} < 1 \wedge \frac{|j|}{h_{\text{tod}}} < 1 \\ 0 & \text{for } \frac{|i|}{h_{\text{doy}}} \geq 1 \vee \frac{|j|}{h_{\text{tod}}} \geq 1 \end{cases} \quad (\text{A.23})$$

where h_{day} is the bandwidth in the *day of year* dimension (in days) and h_{tod} is the bandwidth in the *time of day* dimension (in hours).

References

- P. Bacher, H. Madsen, and H. A. Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83(10):1772–1783, 2009. ISSN 0038092x.
- R. Bird and C. Riordan. Simple solar spectral model for direct and diffuse irradiance on horizontal and tilted planes at the earth’s surface for cloudless atmospheres. Technical report, Solar Energy Research Inst., Golden, CO (USA), 1984.
- R. E. Bird. A simple, solar spectral model for direct-normal and diffuse horizontal irradiance. *Solar Energy*, 32(4):461 – 471, 1984. ISSN 0038-092X. doi: 10.1016/0038-092X(84)90260-3. URL <http://www.sciencedirect.com/science/article/pii/0038092X84902603>.
- W. Cleveland and G. Tiao. Decomposition of seasonal time series: A model for the census x-11 program. *Journal of the American statistical Association*, pages 581–587, 1976.
- J. A. Davies and D. C. McKay. Estimating solar irradiance and components. *Solar Energy*, 29(1):55 – 64, 1982. ISSN 0038-092X. doi: 10.1016/0038-092X(82)90280-8. URL <http://www.sciencedirect.com/science/article/pii/0038092X82902808>.
- DMI. Danish Meteorological Institute, DMI-HIRLAM-S05, 2011. URL http://www.dmi.dk/eng/index/research_and_development/dmi-hirnam-2009.htm.
- DMI. Danish Meteorological Institute, Borgervej. www.borgervej.dk, Feb. 2012.
- J. A. Duffie and W. A. Beckman. *Solar Engineering of Thermal Processes, 3rd Edition*. Wiley, 2006.
- J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.
- M. Geiger, L. Diabaté, L. Ménard, and L. Wald. A web service for controlling the quality of measurements of global solar irradiation. *Solar Energy*, 73(6): 475 – 480, 2002. ISSN 0038-092X. doi: 10.1016/S0038-092X(02)00121-4.

- B. Hansen Sass, N. Woetmann Nielsen, J. U. Jørgensen, B. Amstrup, M. Kmit, and K. S. Mogensen. *The operational DMI-HIRLAM system 2002-version*. DMI, 2002.
- P. Ineichen. Comparison of eight clear sky broadband models against 16 independent data banks. *Solar Energy*, 80(4):468 – 478, 2006. ISSN 0038-092X. doi: 10.1016/j.solener.2005.04.018.
- Isaac and Moradi. Quality control of global solar radiation using sunshine duration hours. *Energy*, 34(1):1 – 6, 2009. ISSN 0360-5442. doi: 10.1016/j.energy.2008.09.006.
- M. Journée and C. Bertrand. Quality control of solar radiation data within the RMIB solar measurements network. *Solar Energy*, 85(1):72 – 86, 2011. ISSN 0038-092X. doi: 10.1016/j.solener.2010.10.021.
- R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.
- R. Koenker. *quantreg: Quantile Regression*, 2011. URL <http://CRAN.R-project.org/package=quantreg>. R package version 4.76.
- R. Mueller, K. Dagestad, P. Ineichen, M. Schroedter-Homscheidt, S. Cros, D. Dumortier, R. Kuhlemann, J. Olseth, G. Piernavieja, C. Reise, L. Wald, and D. Heinemann. Rethinking satellite-based solar irradiance modelling: The solis clear-sky module. *Remote Sensing of Environment*, 91(2):160 – 174, 2004. ISSN 0034-4257. doi: 10.1016/j.rse.2004.02.009.
- J. K. Møller, H. A. Nielsen, and H. Madsen. Time-adaptive quantile regression. *Computational Statistics and Data Analysis*, 52(3):1292–1303, 2008. ISSN 01679473.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL <http://www.R-project.org/>. ISBN 3-900051-07-0.
- C. Rigollier, O. Bauer, and L. Wald. On the clear sky model of the esra - european solar radiation atlas - with respect to the heliosat method. *Solar Energy*, 68(1):33 – 48, 2000. ISSN 0038-092X. doi: 10.1016/S0038-092X(99)00055-9.
- J. Ruiz-Arias, H. Alsamamra, J. Tovar-Pescador, and D. Pozo-Vázquez. Proposal of a regressive model for the hourly diffuse solar radiation under all sky conditions. *Energy Conversion and Management*, 51(5):881–893, 2010. ISSN 01968904. doi: 10.1016/j.enconman.2009.11.024.
- S. Younes, R. Claywell, and T. Muneer. Quality control of solar radiation data: Present status and proposed new approaches. *Energy*, 30(9):1533 – 1549, 2005. ISSN 0360-5442. doi: 10.1016/j.energy.2004.04.031. Measurement and Modelling of Solar Radiation and Daylight- Challenges for the 21st Century.

P A P E R B

Online Short-term Solar Power Forecasting

Authors:

Peder Bacher¹, Henrik Madsen¹, and Henrik Aalborg Nielsen²

Published in:

Solar Energy, 2009, 83(10), pp. 1772-1783.

¹DTU Informatics, Richard Pedersens Plads, Building 321, DK-2800 Lyngby, Denmark

²ENFOR A/S, Lyngsø Allé 3, DK-2970 Hørsholm, Denmark (URL: www.enfor.eu)

Abstract

This paper describes a new approach to online forecasting of power production from PV systems. The method is suited to online forecasting in many applications and in this paper it is used to predict hourly values of solar power for horizons of up to 36 hours. The data used is fifteen-minute observations of solar power from 21 PV systems located on rooftops in a small village in Denmark. The suggested method is a two-stage method where first a statistical normalization of the solar power is obtained using a clear sky model. The clear sky model is found using statistical smoothing techniques. Then forecasts of the normalized solar power are calculated using adaptive linear time series models. Both autoregressive (AR) and AR with exogenous input (ARX) models are evaluated, where the latter takes numerical weather predictions (NWP) as input. The results indicate that for forecasts up to two hours ahead the most important input is the available observations of solar power, while for longer horizons NWP are the most important input. A root mean square error improvement of around 35 % is achieved by the ARX model compared to a proposed reference model.

B.1 Introduction

Efforts to increase the capacity of solar power production in Denmark are concentrating on installing grid connected PV systems on rooftops. The peak power of the installed PV systems is in the range of 1 to 4 kWp, which means that the larger systems will approximately cover the electricity consumption (except heating) of a typical family household in Denmark. The PV systems are connected to the main electricity grid and thus the output from other power production units has to be adjusted in order to balance the total power production. The cost of these adjustments increases as the horizon of the adjustments decreases and thus improved forecasting of solar power will result in an optimized total power production, and in future power production systems where energy storage is implemented, power forecasting is an important factor in optimizing utilization of storage facilities (Koeppel and Korpas, 2006).

The total electricity power production in Denmark is balanced by the energy market Nord Pool, where electricity power is traded on two markets: the main market Elspot and a regulation market Elbas. On Nord Pool the producers release their bids at 12:00 for production each hour the following day, thus the relevant solar power forecasts are updated before 12:00 and consist of hourly values at horizons of 12 to 36 hours. The models in this paper focus on such forecasts, but with the 1-to-11-hour horizons also included.

Interest in forecasting solar power has increased and several recent studies deal with the problem. Many of these consider forecasts of the global irradiance which is essentially the same problem as forecasting solar power. Two approaches are dominant:

- a two-stage approach in which the solar power (or global irradiance) is normalized with a clear sky model in order to form a more stationary time series and such that the classical linear time series methods for forecasting can be used.
- another approach in which neural networks (NNs) with different types of input are used to predict the solar power (or global irradiance) directly.

In a study Chowdhury and Rahman (1987) make sub-hourly forecasts by normalizing with a clear sky model. The solar power is divided into a clear sky component, which is modelled with a physical parametrization of the atmosphere, and a stochastic cloud cover component which is predicted using ARIMA models. Sfetsos and Coonick (2000) use NNs to make one-step predictions of hourly values of global irradiance and compare these with linear time series models that work by predicting clearness indexes. Heinemann et al. (2006) use satellite images for horizons below 6 hours, and in (Lorenz et al., 2007) numerical weather predictions (NWP) for longer horizons, as input to NNs to predict global irradiance. This is transformed into solar power by a simulation model of the PV system. Hocaoglu et al. (2008) investigate feed-forward NNs for one-step predictions of hourly values of global irradiance and compare these with seasonal AR models applied on solar power directly. Cao and Lin (2008) use NNs combined with wavelets to predict next day hourly values of global irradiance. Different types of meteorological observations are used as input to the models; among others the daily mean global irradiance and daily mean cloud cover of the day to be forecasted.

This paper describes a new two-stage method where first the clear sky model approach is used to normalize the solar power and then adaptive linear time series models are applied for prediction. Such models are linear functions between values with a constant time difference, where the model coefficients are estimated by minimizing a weighted residual sum of squares. The coefficients are updated regularly, and newer values are weighted higher than old values, hence the models adapt over time to changing conditions.

Normalization of the solar power is obtained by using a clear sky model which gives an estimate of the solar power in clear (non-overcast) sky at any given point in time. The clear sky model is based on statistical smoothing techniques and quantile regression, and the observed solar power is the only input. The adaptive linear prediction is obtained using recursive least squares (RLS) with

forgetting. It is found that the adaptivity is necessary, since the characteristics of a PV-system are subject to changes due to snow cover, leaves on trees, dirt on the panel, etc., and this has to be taken into account by an online forecasting system.

The data used in the modelling is described in Section B.2. The clear sky model used for normalizing the solar power is defined in Section B.3 followed by Section B.4 where the adaptive time series models used for prediction are identified. In Section B.5 an approach to modelling of the uncertainty in the forecasts is outlined. The evaluation of the models and a discussion of the results are found in Section B.6 and finally the conclusions of the study are drawn in Section B.7.

Nomenclature

p	Solar power	W
p_{cs}	Clear sky solar power	W
τ	Normalized solar power	-
t	Time index	-
k	Forecast horizon index	-
i, j	Miscellaneous indexes	-
p_t	Observation of average solar power	W
$\hat{p}_{t+k t}$	k -step prediction of solar power	W
\hat{p}_t^{cs}	Estimated clear sky solar power	W
$\hat{g}_{i,k}$	i 'th update of NWP of global irradiance	W/m ²
$\hat{g}_{k,t}^{00}$	NWP of global irradiance updated at 00:00	W/m ²
$\hat{g}_{k,t}^{12}$	NWP of global irradiance updated at 12:00	W/m ²
$p_{k,t}^{00}$	Observation of solar power corresponding to $\hat{g}_{k,t}^{00}$	W
$p_{k,t}^{12}$	Observation of solar power corresponding to $\hat{g}_{k,t}^{12}$	W
τ_t	Normalized solar power	-
$\hat{\tau}_{t+k t}$	k -step prediction of normalized solar power	-
$\hat{\tau}_t^{nwp}$	NWPs transformed into normalized solar power	-
x_t	Day of year	-
y_t	Time of day	-
e_{t+k}	k -step prediction error	-
q	Quantile level	-
h	Bandwidth of smoothing kernel	-
λ	Forgetting factor	-

B.2 Data

The data used in this study is observations of solar power from 21 PV systems located in a small village in Jutland, Denmark. The data covers the entire year 2006. Forecasts of global irradiance are provided by the Danish Meteorological Institute using the HIRLAM mesoscale NWP model.

The PV array in each the 21 PV systems is composed of “BP 595” PV modules and the inverters are of the type “BP GCI 1200”. The installed peak power of

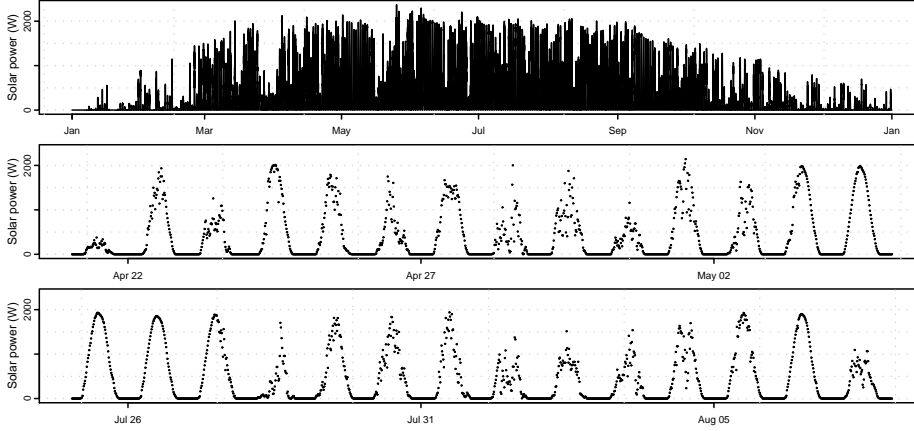


Figure B.1: The observations of average solar power used in the study. Upper plot: The solar power over the entire year 2006. Lower plots: The solar power in two selected periods.

the PV arrays is between 1020 Watt peak and 4080 Watt peak, and the average is 2769 Watt peak. Let $p_{i,t}$ denote the average value of solar power (W) over 15 minutes observed for the i 'th PV system at time t . These observations are used to form the time series

$$\{p_t; t = 1, \dots, N\} \quad (\text{B.1})$$

where

$$p_t = \frac{1}{21} \sum_{i=1}^{21} p_{i,t} . \quad (\text{B.2})$$

This time series is used throughout the modelling. The time series covers the period from 01 January 2006 to 31 December 2006. The observations are fifteen-minute values, ie. $N = 35040$. Plots of $\{p_t\}$ are shown in Figure B.1 for the entire period and for two shorter periods.

The NWP's of global irradiance are given in forecasts of average values for every third hour, and the forecasts are updated at 00:00 and 12:00 each day. The i 'th update of the forecasts is the time series

$$\{\hat{g}_{i,k}, k = 1, \dots, 12\} \quad (\text{B.3})$$

which then covers the forecast horizons up to 36 hours ahead, and is given in (W/m^2).

Time series are resampled to lower sample frequencies by mean values and when the resampled values are used this is noted in the text. In order to synchronize

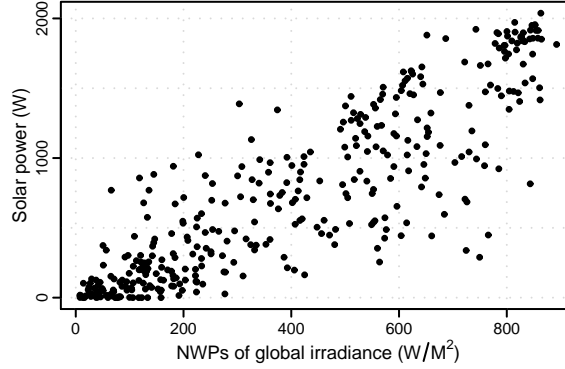


Figure B.2: All three hour interval values of solar power at time of day 10:30 versus the corresponding NWPs of global irradiance with 24 hour horizon. Hence the plot shows observations and predictions of values covering identical time intervals.

data with different sample frequencies, the time point for a given mean value is assigned to the middle of the period that it covers, e.g. the time point of an hourly value of solar power from 10:00 to 11:00 is assigned to 10:30.

As an example of the NWPs of global irradiance Figure B.2 shows values at time of day 10:30 of $\{p_t\}$ resampled to three hour interval values plotted versus the corresponding $\{\hat{g}_{i,k}\}$ values with a 24 hour horizon. Clearly the plot indicates a significant correlation. Hence it is seen that there is information in the NWPs, which can be utilized to forecast the solar power.

B.3 Clear sky model

A clear sky model is usually a model which estimates the global irradiance in clear (non-overcast) sky at any given time. Chowdhury and Rahman (1987) divide the global irradiance into a clear sky component and a cloud cover component by

$$G = G_{cs} \cdot \tau_c \quad (\text{B.4})$$

where G is the global irradiance (W/m^2), and G_{cs} is the clear sky global irradiance (W/m^2). Finally τ_c is the transmissivity of the clouds which they model as a stochastic process using ARIMA models. The clear sky global irradiance is found by

$$G_{cs} = I_0 \cdot \tau_a \quad (\text{B.5})$$

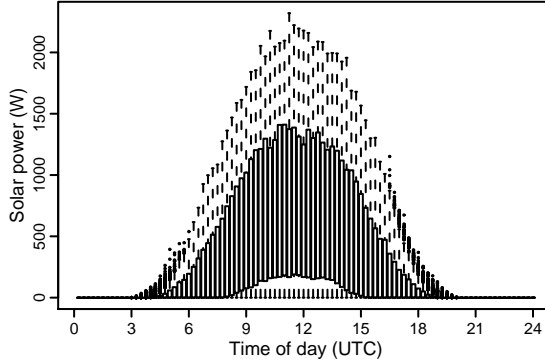


Figure B.3: Modified boxplots of the distribution of the solar power as a function of time of day. The boxplots are calculated with all the fifteen-minutes values of solar power, i.e. covering all of 2006. At each time of the day the box represents the center half of the distribution, from the first to the third quantile. The lower and upper limiting values of the distribution are marked with the ends of the vertical dotted lines, and dots beyond these indicate outliers.

where I_0 is the extraterrestrial irradiance (W/m^2). τ_a is the total sky transmissivity in clear sky which is modelled by atmospheric dependent parametrization.

In this study the same approach is used, but instead of applying the factor on global irradiance it is applied on solar power, i.e.

$$p = p_{cs} \cdot \tau \quad (\text{B.6})$$

where p is the solar power (W) and p_{cs} is the clear sky solar power (W). The factors τ and τ_c are much alike, but since the clear sky model developed in the present study estimates p_{cs} by statistical smoothing techniques rather than using physics, the method is mainly viewed as a statistical normalization technique and τ is referred to as normalized solar power.

The motivation behind the proposed normalization of the solar power with a clear sky model is that the normalized solar power (the ratio of solar power to clear sky solar power) is more stationary than the solar power, so that classical time series models assuming stationarity (Madsen, 2007) can be used for predicting the normalized values. The non-stationarity is illustrated by Figure B.3 where modified boxplots indicate the distribution of solar power p_t as a function of time of day. Clearly a change in the distributions over the day is seen and this non-stationarity must be considered. Figure B.4 shows the same type of plot for the normalized solar power and it is seen that the distributions over the

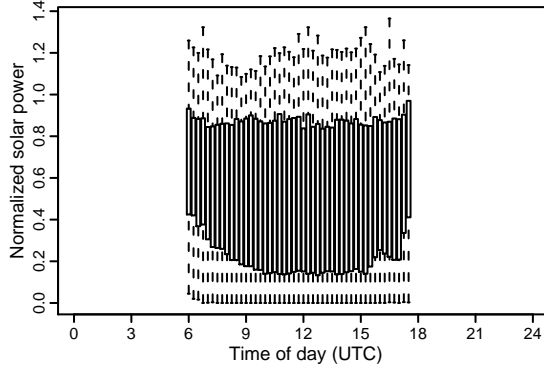


Figure B.4: Modified boxplots of the distribution of the normalized solar power as a function of time of day. The boxplots are calculated with all fifteen-minutes values available, i.e. covering all of 2006.

day are closer to being identical. Thus the effect of the changes over the day is much lower for the normalized solar power than for the solar power.

The clear sky model is defined as

$$p_{cs} = f_{\max}(x, y) \quad (\text{B.7})$$

where p_{cs} is the clear sky solar power (W), x is the day of year and y is the time of day. The function $f_{\max}(\cdot, \cdot)$ is assumed to be a smooth function and thus $f_{\max}(\cdot, \cdot)$ can be estimated as a local maximum (Koenker, 2005). Figure B.5 shows the solar power plotted as a function of x and y , and the estimated clear sky solar power $\hat{f}_{\max}(\cdot, \cdot)$ is shown as a surface in Figure B.6. Due to outliers the weighted quantile regression method outlined in Section B.8 is used to find the local maximum. The $\hat{f}_{\max}(\cdot, \cdot)$ is then used to form the output of the clear sky model as the time series

$$\{\hat{p}_t^{cs}, t = 1, \dots, N\}, \quad (\text{B.8})$$

where \hat{p}_t^{cs} is the estimated clear sky solar power (W) at time t , and $N = 35040$. The normalized solar power is now defined as

$$\tau_t = \frac{p_t}{\hat{p}_t^{cs}} \quad (\text{B.9})$$

and this is used to form time series of normalized solar power

$$\{\tau_t, t = 1, \dots, 35040\}. \quad (\text{B.10})$$

For each (x_t, y_t) corresponding to the solar power observation p_t , weighted quantile regression estimates the q quantile by a Gaussian two dimensional smoothing kernel, defined in Section B.8. The smoothing kernel is used to form the

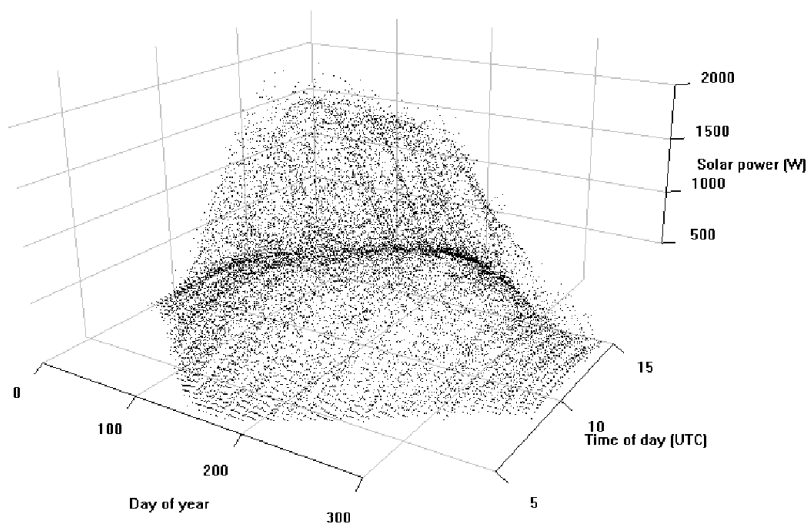


Figure B.5: The solar power as a function of the day of year, and the time of day. Note that only positive values of solar power are plotted.

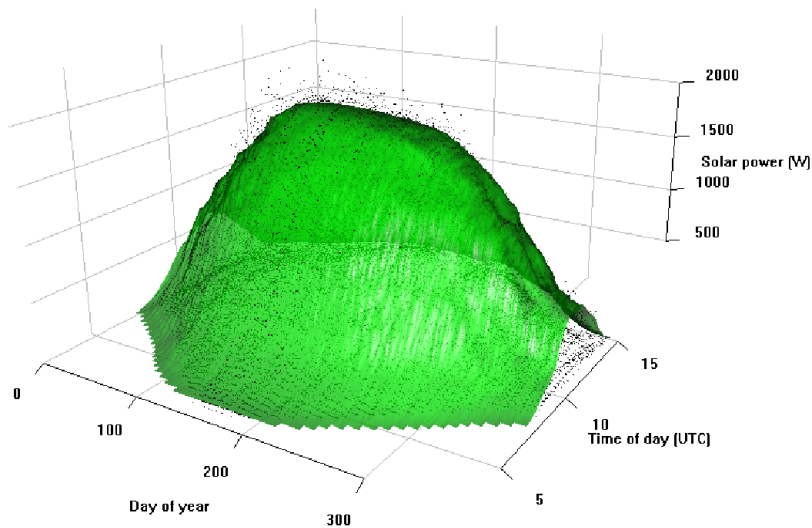


Figure B.6: The estimated clear sky solar power shown as a surface. The solar power is shown as points.

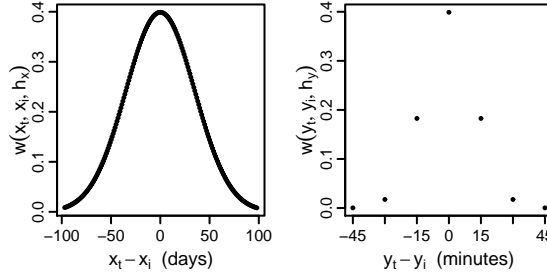


Figure B.7: The one dimensional smoothing kernels used. Left plot is the kernel in the day of year (x) dimension. Right plot is the kernel in the time of day (y) dimension. They are multiplied to form the applied two dimensional smoothing kernel.

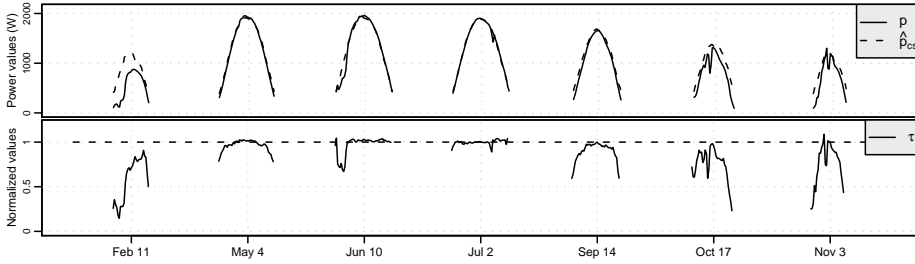


Figure B.8: The result of the normalization for selected clear sky days over the year. The time-axis ticks refer to midday points, i.e. at 12:00. The upper plot shows the solar power p and the estimated clear sky solar power \hat{p}_{cs} . The lower plot shows the normalized solar power τ .

weights applied in the quantile regression. As seen in Figure B.7, which shows the smoothing kernel used, the weights in the day of year dimension $w(x_t, x_i, h_x)$, are decreasing as the absolute time differences are increasing. Similarly for the weights in the time of day dimension $w(y_t, y_i, h_y)$. The applied weights are finally found by multiplying the weights from the two dimensions. The choice of the quantile level q to be estimated and the bandwidth in each dimension, h_x and h_y , is based on a visual inspection of the results. A level of $q = 0.85$ was used since this gives $\tau_t \approx 1$ for days with clear sky all day, as seen in Figure B.8. The plot for days with varying cloud cover in Figure B.9 show that estimates where $\tau_t > 1$ occur. These are ascribed to reflections from clouds and varying level of water vapour in the atmosphere. Future work should elaborate on the inclusion of such effects in the clear sky model.

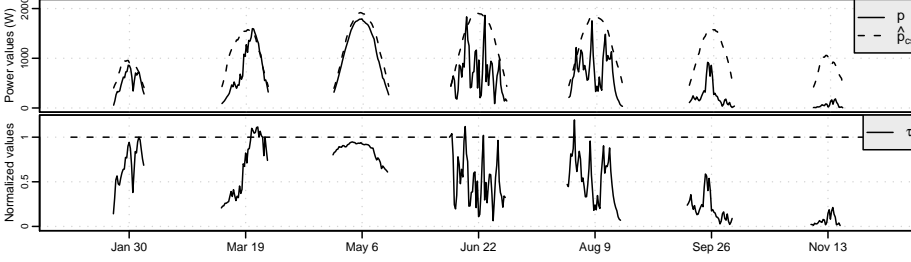


Figure B.9: The result of the normalization for days evenly distributed over the year. The time-axis ticks refer to midday points, i.e. at 12:00. The upper plot shows the solar power p and the estimated clear sky solar power \hat{p}_{cs} . The lower plot shows the normalized solar power τ .

For small \hat{p}_t^{cs} values the error of τ_t is naturally increasing and at nighttime the error is infinite. Therefore all values of \hat{p}_t^{cs} where

$$\frac{\hat{p}_t^{cs}}{\max(\{\hat{p}_t^{cs}\})} < 0.2 \quad (\text{B.11})$$

are removed from $\{\tau_t\}$. The function $\max(\{\hat{p}_t^{cs}\})$ gives the maximum value in $\{\hat{p}_t^{cs}\}$.

The estimates of clear sky solar power are best in the summer period. The bad estimates in winter periods are caused by the sparse number of clear sky observations. It should also be possible to improve the normalization toward dusk and dawn, and thus lower the limit where values in $\{\hat{p}_{cs}\}$ are removed, either by refining the modelling method or by including more explanatory variables such as e.g. air mass.

Finally it is noted that the deterministic changes of solar power are really caused by the geometric relation between the earth and the sun, which can be represented in the current problem by the sun elevation as x and sun azimuth as y . The clear sky solar power was also modelled in the space spanned by these two variables, by applying the same statistical methods as for the space spanned by day of year and time of day. The result was not satisfactory, i.e. the estimated clear sky solar power was less accurate, probably because neighboring values in this space are not necessarily close in time and thus changes in the surroundings to the PV system blurred the estimates.

B.4 Prediction models

Adaptive linear time series models (Madsen, 2007) are applied to predict future values of the normalized solar power τ_t . The inputs are: lagged observations of τ_t and transformed NWP $\hat{\tau}_t^{\text{nwp}}$. Three types of models are identified:

- a model which has only lagged observations of τ_t as input. This is an autoregressive (AR) model and it is referred to as the *AR* model.
- a model with only $\hat{\tau}_t^{\text{nwp}}$ as input. This is referred to as the *LM_{nwp}* model.
- a model with both types of input. This is an autoregressive with exogenous input (ARX) model and it is referred to as the *ARX* model.

The best model of each type is identified by using the autocorrelation function (ACF).

B.4.1 Transformation of NWPs into predictions of normalized solar power

In order to use the NWPs of global irradiation $\hat{g}_{i,k}$ as input to the prediction models, these are transformed into $\hat{\tau}_t^{\text{nwp}}$ which are meteorological based hourly predictions of τ_t . This is done by first transforming $\hat{g}_{i,k}$ into solar power predictions and then transforming these by the clear sky model. The time series $\{\hat{g}_{i,k}\}$, defined in (B.3), holds the i 'th NWP forecast of three hour interval values, and was updated at

$$\text{time}_i = t_0 + (i - 1) \cdot 12\text{h} \quad (\text{B.12})$$

where $t_0 = 2006-01-01$ 00:00. Thus the time series with sample period of one day

$$\{\hat{g}_{k,t}^{00}, t = 1, \dots, 364\} = \{\hat{g}_{i,k}, i = 1, 3, \dots, 727\}, \quad (\text{B.13})$$

consist of all the NWPs updated at time of day 00:00 at horizon k , i.e. the superscript “00” forms part of the name of the variable. Similarly the time series

$$\{\hat{g}_{k,t}^{12}, t = 1, \dots, 364\} = \{\hat{g}_{i,k}, i = 2, 4, \dots, 728\}, \quad (\text{B.14})$$

consist of all the NWPs updated at time of day 12:00. The corresponding time series of solar power covering the identical time intervals are respectively

$$\begin{aligned} \{p_{k,t}^{00}, t = 1, \dots, 364\} = \\ \{p_t, t = k, (1 \cdot 8 + k), \dots, (363 \cdot 8 + k)\} \end{aligned} \quad (\text{B.15})$$

and

$$\{p_{k,t}^{12}, t = 1, \dots, 364\} = \{p_t, t = k + 4, (1 \cdot 8 + k + 4), \dots, (363 \cdot 8 + k + 4)\}, \quad (\text{B.16})$$

where $\{p_t\}$ has been resampled to three hour interval values. The NWP are modelled into solar power predictions by the adaptive linear model

$$\hat{p}_{k,t}^{00} = \beta_t + \alpha_t \hat{g}_{k,t}^{00} + e_t, \quad (\text{B.17})$$

note that the hat above the variable indicates that these values are predictions (estimates) of the solar power. A similar model is made for the NWP updates at time of day 12:00 giving $\hat{p}_{k,t}^{12}$. The interpretation of the coefficients β_t and α_t is not further elaborated here, but it is noted that they are time dependent in order to account for the effects of changing conditions over time, e.g. the changing geometric relation between the earth and the sun, dirt on the solar panel. This adaptivity is obtained by fitting the model with k -step recursive least squares (RLS) with forgetting, which is described in Section B.9. In order to use the RLS algorithm, $p_{k,t}^{00}$ has to be lagged depending on k . Each RLS estimation is optimized by choosing the value of the forgetting factor λ from 0.9, 0.905, \dots , 1 that minimizes the root mean square error (*RMSE*).

The last steps in the transformation of the NWPs is to normalize $\hat{p}_{k,t}^{00}$ and $\hat{p}_{k,t}^{12}$ with the clear sky model, and resample up to hourly values by linear interpolation. Finally the time series

$$\{\hat{\tau}_t^{\text{nwp}}, t = 1, \dots, 8760\} \quad (\text{B.18})$$

of the NWPs of global irradiance transformed into predictions of normalized solar power is formed, and this is used as input to the *ARX* prediction models as described in the following. More details can be found in (Bacher, 2008).

B.4.2 AR model identification

To investigate the time dependency in $\{\tau_t\}$, i.e. dependency between values with a constant time difference, the ACF is calculated and plotted in Figure B.10. Clearly an AR(1) component is indicated by the exponential decaying pattern of the first few lags and a seasonal diurnal AR component by the exponential decaying peaks at lag = 24, 48, \dots . By considering only first-order terms this leads to the 1-step AR model

$$\tau_{t+1} = m + a_1 \tau_t + a_2 \tau_{t-23} + e_{t+1}. \quad (\text{B.19})$$

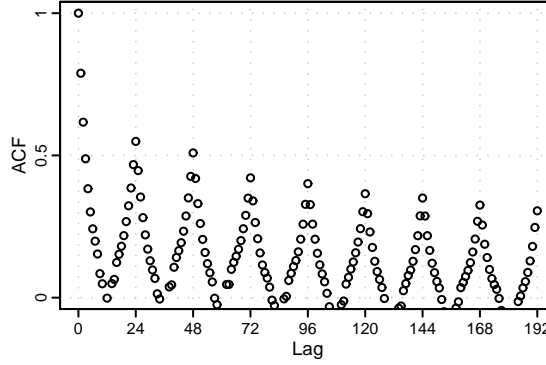


Figure B.10: ACF of the time series of normalized solar power $\{\tau_t\}$.

And a reasonable 2-step AR model is

$$\tau_{t+2} = m + a_1\tau_t + a_2\tau_{t-22} + e_{t+2} . \quad (\text{B.20})$$

Note that here the 1-step lag cannot be used, since this is τ_{t+1} i.e. a future value, and thus the latest observed value τ_t is included instead. Formulated as a k -step AR model

$$\tau_{t+k} = m + a_1\tau_t + a_2\tau_{t-s(k)} + e_{t+k} \quad (\text{B.21})$$

$$s(k) = 24 + k \bmod 24 \quad (\text{B.22})$$

where the function $s(k)$ ensures that the latest observation of the diurnal component is included. This is needed, since for $k = 25$ the diurnal 24 hour AR component cannot be used and instead the 48 hour AR component is used. This model is referred to as the *AR* model.

Figure B.11 shows the ACF of $\{e_{t+k}\}$, which is the time series of the errors in the model for horizon k , for six selected horizons after fitting the *AR* model with RLS, which is described in Section B.9. The vertical black lines indicate which lags are included in the model. For $k = 1$ the correlation of the AR(1) component is removed very well and the diurnal AR component has also been decreased considerably. There is high correlation left at lag = 24, 48, ... This can most likely be ascribed to systematic errors caused by non-stationarity effects left in $\{\tau_t\}$, and it indicates that the clear sky model normalization can be further optimized. For $k = 2$ and 3 the grayed points show the lags that cannot be included in the model and the high correlation of these lags indicate that information is not exploited. The *AR* model was extended with higher order AR and diurnal AR terms without any further improvement in performance, see (Bacher, 2008).

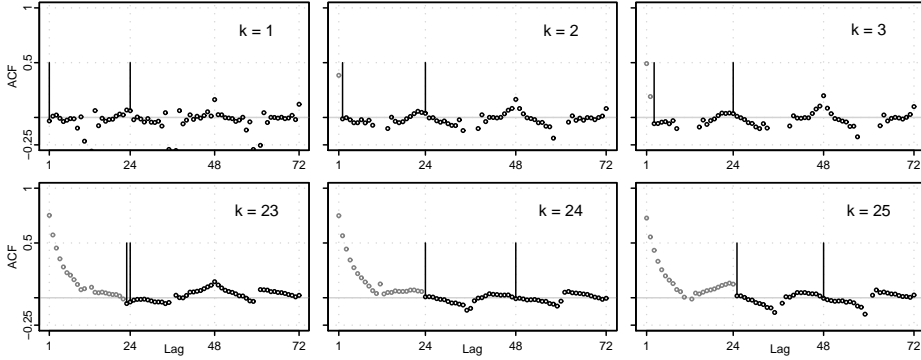


Figure B.11: ACF of the time series of errors $\{e_{t+k}\}$ for selected horizons k of the AR model. The vertical bars indicate the lags included in each of the models, and the grayed points show the lags which cannot be included in the model.

B.4.3 LM_{nwp} model identification

The model using only NWP as input

$$\tau_{t+k} = m + b_1 \hat{\tau}_{t+k|t}^{\text{nwp}} + e_{t+k} \quad (\text{B.23})$$

is referred to as LM_{nwp} . It is fitted using RLS and the ACF of $\{e_{t+k}\}$ is shown in Figure B.12 for two horizons. For $k = 1$ clearly correlation is left from an AR(1) component, but as seen for both horizons the actual NWP input removes diurnal correlation very well.

B.4.4 ARX model identification

The model using both lagged observations of τ_t and NWP as input is an ARX model. The LM_{nwp} revealed an exponential decaying ACF for short horizons and thus an AR(1) term is clearly needed, whereas adding the diurnal AR component has only a small effect. The results show that in fact the diurnal AR component can be left out, but it is retained since this clarifies that no improvement is achieved by adding it, this is showed later. The model

$$\tau_{t+k} = m + a_1 \tau_t + a_2 \tau_{t-s(k)} + b_1 \hat{\tau}_{t+k|t}^{\text{nwp}} + e_{t+k}, \quad (\text{B.24})$$

is referred to as the ARX model. The model is fitted using RLS and the ACF of $\{e_{t+k}\}$ is plotted in Figure B.13. It is seen that the AR(1) component removes

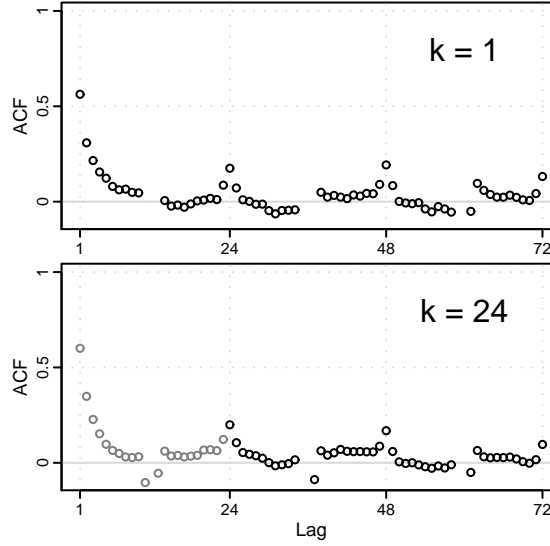


Figure B.12: ACF of the time series of errors $\{e_{t+k}\}$ at horizon $k = 1$ and $k = 24$ of the LM_{nwp} model. The grayed points show the lags which cannot be included in the model.

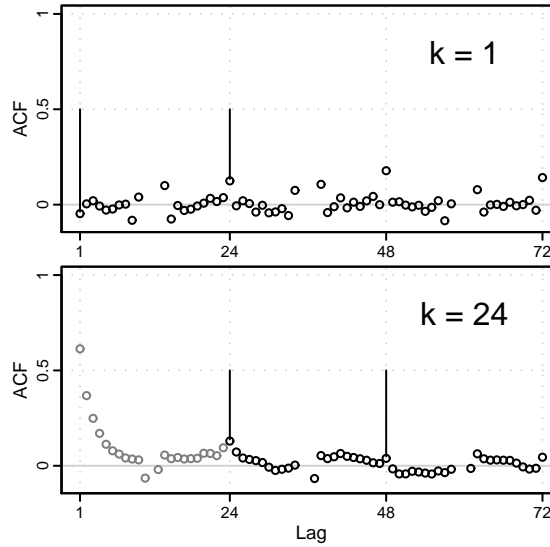


Figure B.13: ACF of the time series of errors $\{e_{t+k}\}$ at horizons $k = 1$ and $k = 24$ of the ARX model. The vertical bars indicate the lags included in each of the models, and the grayed points show the lags which cannot be included in the model.

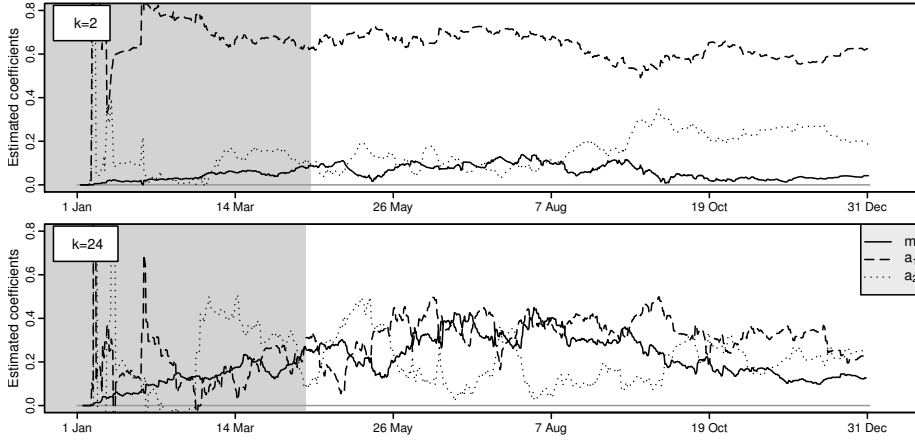


Figure B.14: The online estimates of the coefficients in the AR model as a function of time. Two selected horizons are shown. The grayed period in the beginning marks the burn-in period.

the correlation for the short horizons very well. The *ARX* was extended with higher order AR and diurnal AR terms without any further improvements in performance.

B.4.5 Adaptive coefficient estimates

The plots in Figure B.14 show the online coefficient estimates for the *AR* model, where a value of $\lambda = 0.995$ is used since this is the value that minimizes the $RMSE_k$ best for all horizons in the current setting. Clearly the values of the coefficient estimates change over time and this indicates that the adaptivity is needed to make an optimal model for online forecasting.

B.5 Uncertainty modelling

Extending the solar power forecasts, from predicting a single value (a point forecast) to predicting a distribution increases their usefulness. This can be achieved by modelling the uncertainties of the solar power forecasts and a simple approach is outlined here. The classical way of assuming normal distribution of the errors will in this case not be appropriate since the distribution of the errors has finite limits. Instead, quantile regression is used, inspired by Møller et al.

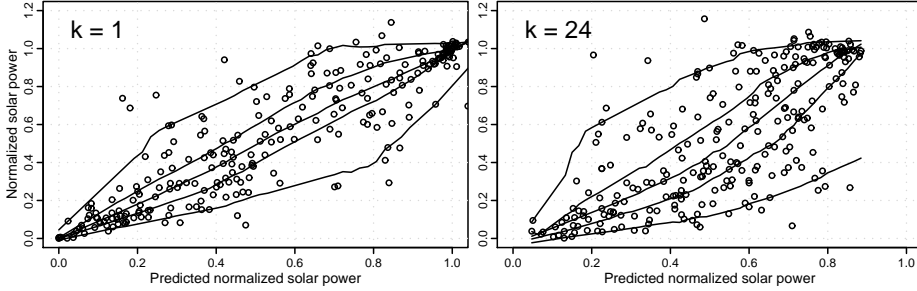


Figure B.15: Normalized solar power versus the predicted normalized solar power at horizons $k = 1$ and $k = 24$. The predictions are made with the *ARX* model. The lines are estimates of the 0.05, 0.25, 0.50, 0.75 and 0.95 quantiles of $f_{\tau}(\hat{\tau})$.

(2008) where it is applied to wind power forecasts. Plots of $\{\tau_t\}$ versus $\{\hat{\tau}_t\}$ for a given horizon reveal that the uncertainties depend on the level of $\hat{\tau}$. Figure B.15 shows such plots for horizons $k = 1$ and $k = 24$. The lines in the plot are estimates of the 0.05, 0.25, 0.50, 0.75 and 0.95 quantiles of the probability distribution function of τ as a function of $\hat{\tau}$. The weighted quantile regression with a one dimensional kernel smoother, described in Section B.8, is used.

Figure B.15 illustrates that the uncertainties are lower for $\hat{\tau}$ close to 0 and 1, than for the mid-range values around 0.5. Thus forecasts of values toward overcast or clear sky have less uncertainty than forecasts of a partly overcast sky, which agrees with results by Lorenz et al. (2007). Further work should extend the uncertainty model to include NWP as input.

B.6 Evaluation

The methods used for evaluating the prediction models are inspired by Madsen et al. (2005) where a framework for evaluation of wind power forecasting is suggested. The RLS fitting of the prediction models does not use any degrees of freedom and the dataset is therefore not divided into a training set and a test set. It is, however, noted that the clear sky model and the optimization of λ does use the entire dataset, and thus the results can be a little optimistic. The values in the burn-in period are not used in calculating the error measures. In Figure B.14 the burn-in periods for the *AR* model are shown.

B.6.1 Error measures

The k -step prediction error is

$$e_{t+k} = p_{t+k} - \hat{p}_{t+k|t} \quad (\text{B.25})$$

The Root Mean Square Error for the k 'th horizon is

$$RMSE_k = \left(\frac{1}{N} \sum_{t=1}^N e_{t+k}^2 \right)^{\frac{1}{2}}. \quad (\text{B.26})$$

The $RMSE_k$ is used as the main evaluation criterion (EC) for the performance of the models. The Normalized Root Mean Square Error is found by

$$NRMSE_k = \frac{RMSE_k}{p_{\text{norm}}} \quad (\text{B.27})$$

where either

$$p_{\text{norm}} = \bar{p} = \frac{1}{N} \sum_{t=1}^N p_t. \quad (\text{B.28})$$

or p_{norm} is the average peak power of the 21 PV systems.

The mean value of the $RMSE_k$ for a range of horizons

$$\overline{RMSE}_{k_s, k_e} = \frac{1}{k_e - k_s + 1} \sum_{k=k_s}^{k_e} RMSE_k \quad (\text{B.29})$$

is used as a summary error measure. When comparing the performance of two models the improvement

$$I_{\text{EC}} = 100 \cdot \frac{EC_{\text{ref}} - EC}{EC_{\text{ref}}} (\%) \quad (\text{B.30})$$

is used, where EC is the considered evaluation criterion.

B.6.2 Reference model

To compare the performance of prediction models, and especially when making comparisons between different studies, a common reference model is essential. A reference model for solar power is here proposed as the best performing naive predictor for the given horizon. Three naive predictors of solar power are found to be relevant. Persistence

$$p_{t+k} = p_t + e_{t+k}, \quad (\text{B.31})$$

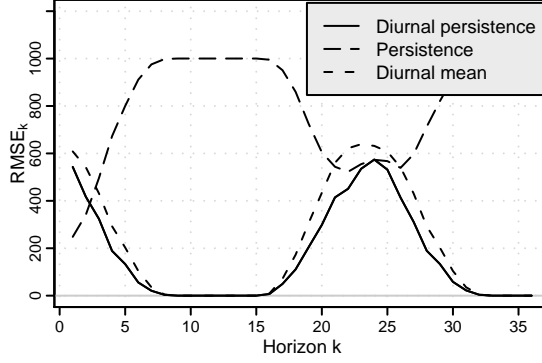


Figure B.16: $RMSE_k$ for the three naive predictors used in the Reference model.

diurnal persistence

$$p_{t+k} = p_{t-s(k)} + e_{t+k} \quad (\text{B.32})$$

$$s(k) = f_{\text{spd}} + k \bmod f_{\text{spd}} \quad (\text{B.33})$$

where $s(k)$ ensures that the latest diurnal observation is used and f_{spd} is the sample frequency in number of samples per day, and diurnal mean

$$p_{t+k} = \frac{1}{n} \sum_{i=1}^n p_{t-s(k,i)} + e_{t+k} \quad (\text{B.34})$$

$$s(k, i) = i \cdot f_{\text{spd}} + k \bmod f_{\text{spd}} \quad (\text{B.35})$$

which is the mean of solar power of the last n observations at the time of day of $t + k$. The value of n is chosen such that all past samples are included.

Figure B.16 shows the $RMSE_k$ for each of the three naive predictors. It is seen that for $k \leq 2$ the persistence predictor is the best while the best for $k > 2$ is the diurnal persistence predictor. This model is referred to as the *Reference* model.

B.6.3 Results

Examples of solar power forecasts made with the *ARX* model are shown in Figure B.17 for short horizons and in Figure B.18 for next day horizons. It is found that the forecasted solar power generally follows the main level of the solar power, but the fluctuations caused by sudden changes in cloud cover are not fully described by the model.

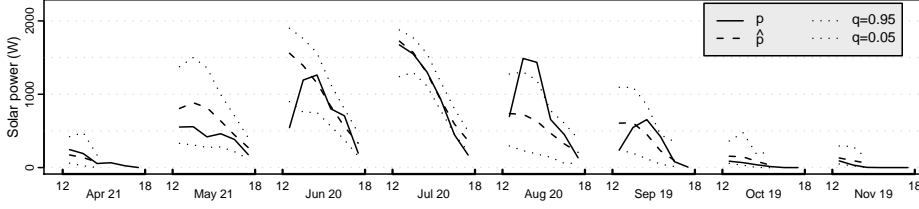


Figure B.17: Forecasts of solar power at short horizons $k = 1, \dots, 6$ made with the *ARX* model.

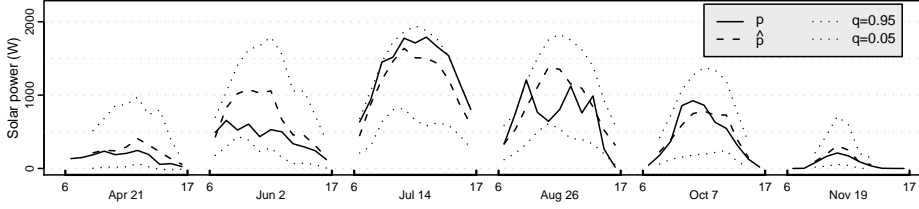


Figure B.18: Forecasts of solar power at next day horizons $k = 19, \dots, 29$ made with the *ARX* model.

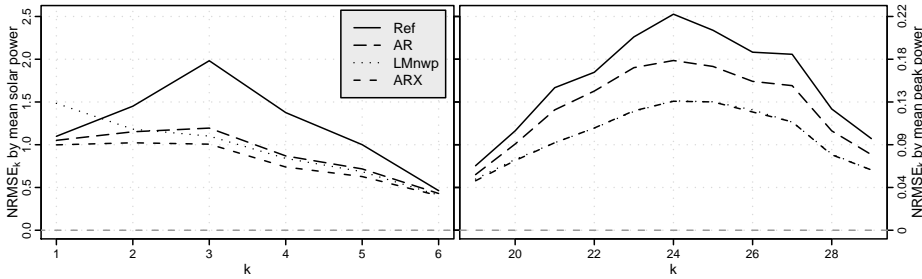


Figure B.19: The $NRMSE_k$ for each of the three models and the Reference model. The left plot show the short horizons and the right the next day horizons. The left scale show $RMSE_k$ normalized by the mean solar power $\bar{p} = 248$ W/h and the right scale show $RMSE_k$ normalized by 2769W, which is the mean peak power of the 21 PV systems.

Models	$I_{\overline{RMSE}_{1,6}}$	$I_{\overline{RMSE}_{19,29}}$
<i>AR</i> over <i>Reference</i>	27%	17%
LM_{nwp} over <i>Reference</i>	25%	36%
<i>ARX</i> over <i>Reference</i>	35%	36%
LM_{nwp} over <i>AR</i>	-2%	23%
<i>ARX</i> over <i>AR</i>	12%	23%
<i>ARX</i> over LM_{nwp}	13%	1%

Table B.1: Summary error measures of improvements compared to the *Reference* model for short horizons $k = 1, \dots, 6$ and next day horizons $k = 19, \dots, 29$.

The $NRMSE_k$ is plotted for each model in Figure B.19. Clearly the performance is increasing from the *Reference* model to the *AR* model and further to the *ARX* model. The differences from using either the solar power or the NWP, or both, as input become apparent from these results.

At $k = 1$ the *AR* model that only uses solar power as input is better than the LM_{nwp} which only uses NWP as input, but at $k = 2, \dots, 6$ the LM_{nwp} is better, though only slightly. This indicates that for making forecasts of horizons shorter than 2 hours, solar power is the most important input, whereas for 2 to 6 hours horizons, forecasting systems using either solar power or NWP can perform almost equally. The *ARX* model using both types of input does have an increased performance at all $k = 1, \dots, 6$ and thus combining the two types of input is found to be the superior approach.

For $k = 19, \dots, 29$, which are the next day horizons, very clearly the LM_{nwp} model and the *ARX* model perform better than the *AR* model. Since the LM_{nwp} model and the *ARX* model perform almost equally, it is seen that no improvement is achieved from adding the solar power as input, and thus using only the NWP as input is found to be adequate for next day horizons.

A summary of the improvement in performance is calculated using (B.29) and (B.30). The improvements compared to the *Reference* model are calculated for the four models by $I_{\overline{RMSE}_{1,6}}$ for short horizons and $I_{\overline{RMSE}_{19,29}}$ for next day horizons. The results are shown in Table B.1. These results naturally show the same as stated above, though the difference at $k = 1$ from *AR* to LM_{nwp} cannot be seen. These results show that a RMSE improvement of around 35 % over the *Reference* model can be achieved by using the *ARX* model.

B.7 Conclusions

Inspired by previous studies, the present method for solar power forecasting has been developed from scratch. A new approach to clear sky modelling with statistical smoothing techniques has been proposed, and an adaptive prediction model based on RLS makes a solid framework allowing for further refinements and model extensions e.g. by including NWP's of temperature as input. The adaptivity of the method makes it suited to online forecasting and ensures comprehension of changing conditions of the PV system and its surroundings. Furthermore the RLS algorithm is not computer intensive, which makes updating of forecasts fast. The clear sky model used to normalize the solar power delivers a useful result, but can be improved, especially for the estimates toward dawn and dusk, by using polynomial-based kernel regression. A procedure based on quantile regression is suggested for calculating the varying intervals of the uncertainty of the solar power predictions and the results agree with other studies. The best performing prediction model is an ARX model where both solar power observations and NWP's are used as input. The results indicate that for horizons below 2 hours solar power is the most important input, but for next day horizons no considerable improvement is achieved from using available values of solar power, so it is adequate just to use NWP's as input. Thus, depending on the application of the forecasting system using only either of the inputs can be considered, and a lower limit of the latency, at which solar power observations are needed for the forecasting system, can be different. Finally it is noted that a comparison to other online solar power forecasting methods, e.g. (Lorenz et al., 2007) and (Hocaoglu et al., 2008), has not been carried out, but that such a study would be informative in order to describe strengths and accuracy of the different proposed methods.

B.8 Weighted quantile regression

The solar power time series $\{p_t, t = 1, \dots, N\}$ is the realization of a stochastic process $\{P_t, t = 1, \dots, N\}$. The estimated clear sky solar power at time t is \hat{p}_t^{cs} and it is found as the q quantile of f_{P_t} , the probability distribution function of P_t . The problem is reduced to estimating \hat{p}_t^{cs} as a local constant for each (x_t, y_t) , where x is the day of year and y the time of day. This is done by weighted quantile regression in which the loss function is

$$\rho(q, \epsilon_i) = \begin{cases} q\epsilon_i & , \quad \epsilon_i \geq 0 \\ (1-q)\epsilon_i & , \quad \epsilon_i < 0 \end{cases} \quad (\text{B.36})$$

where $\epsilon_i = p_i - \hat{p}_t^{\text{cs}}$. The fitting of \hat{p}_t^{cs} is then done by

$$\hat{p}_t^{\text{cs}} = \arg \min_{\hat{p}_t^{\text{cs}}} \sum_{i=1}^N k(x_t, y_t, x_i, y_i) \cdot \rho(q, \epsilon_i). \quad (\text{B.37})$$

where

$$k(x_t, y_t, x_i, y_i) = \frac{w(x_t, x_i, h_x) \cdot w(y_t, y_i, h_y)}{\sum_{i=1}^N w(x_t, x_i, h_x) \cdot w(y_t, y_i, h_y)} \quad (\text{B.38})$$

is the two dimensional multiplicative kernel function which weights the observations locally to (x_t, y_t) , (Hastie and Tibshirani, 1993). Details of the minimization are found in (Koenker, 2005). In each dimension a Gaussian kernel is used

$$w(x_t, x_i, h_x) = f_{\text{std}}\left(\frac{|x_t - x_i|}{h_x}\right) \quad (\text{B.39})$$

where f_{std} is the standard normal probability density function. A similar kernel function is used in the y dimension, and the final two dimensional kernel is found by multiplying the two kernels as shown in (B.37).

B.9 Recursive least squares

Fitting of the prediction models is done using k -step recursive least squares (RLS) with forgetting, which is described in the following using the *ARX* model

$$\tau_{t+k} = m + a_1 \tau_t + a_2 \tau_{t-s(k)} + b_1 \hat{\tau}_{t+k|t}^{\text{nwp}} + e_{t+k}, \quad (\text{B.40})$$

as an example. The regressor at time t is

$$\mathbf{X}_t^T = (1, \tau_t, \tau_{t-s(k)}, \hat{\tau}_{t+k|t}^{\text{nwp}}), \quad (\text{B.41})$$

the parameter vector is

$$\theta^T = (m, a_1, a_2, b_1), \quad (\text{B.42})$$

and the dependent variable

$$Y_t = \tau_t. \quad (\text{B.43})$$

Hence the model can be written as

$$Y_t = \mathbf{X}_t^T \theta + e_t. \quad (\text{B.44})$$

The estimates of the parameters at t are found such that

$$\hat{\theta}_t = \arg \min_{\theta} S_t(\theta), \quad (\text{B.45})$$

where the loss function is

$$S_t(\theta) = \sum_{s=1}^t \lambda^{t-s} (Y_s - \mathbf{X}_s^T \theta)^2. \quad (\text{B.46})$$

This provides weighted least squares with exponential forgetting. The solution at time t leads to

$$\hat{\theta}_t = \mathbf{R}_t^{-1} \mathbf{h}_t, \quad (\text{B.47})$$

see (Madsen, 2007), where

$$\mathbf{R}_t = \sum_{s=1}^t \lambda^{t-s} \mathbf{X}_s \mathbf{X}_s^T, \quad \mathbf{h}_t = \sum_{s=1}^t \lambda^{t-s} \mathbf{X}_s Y_s. \quad (\text{B.48})$$

The k -step RLS-algorithm with exponential forgetting is then

$$\mathbf{R}_t = \lambda \mathbf{R}_{t-1} + \mathbf{X}_{t-k} \mathbf{X}_{t-k}^T \quad (\text{B.49})$$

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \mathbf{R}_t^{-1} \mathbf{X}_{t-k} (Y_t - \mathbf{X}_{t-k}^T \hat{\theta}_{t-1}) \quad (\text{B.50})$$

and the k -step prediction at t is

$$\hat{Y}_{t+k} = \mathbf{X}_t^T \hat{\theta}_t. \quad (\text{B.51})$$

References

- P. Bacher. Short-term solar power forecasting. Master's thesis, Technical University of Denmark, 2008. IMM-M.Sc.-2008-13.
- J. Cao and X. Lin. Study of hourly and daily solar irradiation forecast using diagonal recurrent wavelet neural networks. *Energy Conversion and Management*, 49(6):1396–1406, 2008. ISSN 01968904.
- B. Chowdhury and S. Rahman. Forecasting sub-hourly solar irradiance for prediction of photovoltaic output. In *IEEE Photovoltaic Specialists Conference, 19th, New Orleans, LA, May 4-8, 1987, Proceedings (A88-34226 13-44)*. New York, Institute of Electrical and Electronics Engineers, Inc., 1987, p. 171-176., pages 171–176, 1987.
- T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993. ISSN 00359246.
- D. Heinemann, E. Lorenz, and M. Girodo. Forecasting of solar radiation. In E. Dunlop, L. Wald, and M. Suri, editors, *Solar Resource Management for Electricity Generation from Local Level to Global Scale*, pages 83–94, New York, 2006. Nova Science Publishers.

- F. O. Hocaoglu, Ö. N. Gerek, and M. Kurban. Hourly solar radiation forecasting using optimal coefficient 2-d linear filters and feed-forward neural networks. *Solar Energy*, 82(8):714–726, 2008. ISSN 0038092x.
- R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.
- G. Koeppel and M. Korpas. Using storage devices for compensating uncertainties caused by non-dispatchable generators. *2006 International Conference on Probabilistic Methods Applied to Power Systems*, pages 1–8, 2006.
- E. Lorenz, D. Heinemann, H. Wickramaratne, H. Beyer, and S. Bofinger. Forecast of ensemble power production by grid-connected pv systems. In *Proc. 20th European PV Conference, September 3-7, 2007, Milano*, 2007.
- H. Madsen. *Time Series Analysis*. Chapman & Hall, 2007.
- H. Madsen, P. Pinson, G. Kariniotakis, H. A. Nielsen, and T. S. Nielsen. Standardizing the performance evaluation of shortterm wind power prediction models. *Wind Engineering*, 29(6):475, 2005. ISSN 0309524x.
- J. K. Møller, H. A. Nielsen, and H. Madsen. Time-adaptive quantile regression. *Computational Statistics and Data Analysis*, 52(3):1292–1303, 2008. ISSN 01679473.
- A. Sfetsos and A. Coonick. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. *Solar Energy*, 68(2):169–178, 2000. ISSN 0038092x.

P A P E R C

Online Short-term Solar Power Forecasting

Authors:

Peder Bacher¹, Henrik Madsen¹, and Henrik Aalborg Nielsen²

In proceedings:

1st International Workshop on the Integration of Solar Power into Power Systems. Aarhus, 2011.

¹DTU Informatics, Richard Pedersens Plads, Building 321, DK-2800 Lyngby, Denmark

²ENFOR A/S, Lyngsø Allé 3, DK-2970 Hørsholm, Denmark (URL: www.enfor.eu)

Abstract

This paper describes two methods for online forecasting of power production from PV systems. The methods are suited for online forecasting in many applications and in this paper they are used to predict hourly values of solar power for horizons up to 32 hours. The data used is hourly observations of solar power from a single PV system located on a rooftop in a small village in Denmark. One approach is a two-stage method in which a statistical normalization of the solar power is obtained using a clear sky model. The clear sky model is found using statistical smoothing techniques, which ensure that local phenomena are directly modelled from data, as opposed to applying a deterministically derived clear sky model. In the second stage forecasts of the normalized solar power are calculated using adaptive linear time series models. A second approach is to apply conditional parametric models with both autoregressive input and NWP's exogenous input. The results indicate that for forecasts up to two hours ahead the most important input is the available observations of solar power, while for longer horizons NWP's are the most important input. A root mean square error improvement over a persistence model around 40 % is achieved for 1 and 2 hour horizons and around 35 % for longer horizons.

C.1 Introduction

The increasing installed solar power capacity rises the challenges of grid integration. The need for efficient forecasting methods is evident and the research activities within the topic is increasing, see for example (Sfetsos and Coonick, 2000), (Hocaoglu et al., 2008), (Lorenz et al., 2009), and (Ji and Chee, 2011). In this paper methods for online forecasting are presented. The methods are suited for forecasting of solar power for different systems and here they are applied to forecast the power production of a single 4 kW-peak PV-system installed on a rooftop of a single family house. Due to the fluctuating nature of solar power such forecasts are essential for optimal grid integration and will be essential for solar power smart grid technology. The applications include energy trading for large solar power producers, and diurnal peak-shaving and cost optimization for smaller systems with storage capacity in battery packs (e.g. provided in an electrical car). Two approaches are considered. One is based on a two-stages approach: first the systematic dependency of the position of the sun relative to the PV panel are removed with a clear sky model, and secondly the resulting process is forecasted with time-adaptive linear time series methods. The clear sky model is calculated with non-linear statistical techniques, which will also model the local conditions, such as e.g. shadows from elements in the surrounding

environment and snow cover. In the second approach numerical weather predictions (NWP) are used as input to conditional parametric non-linear models (Nielsen et al., 2002) to forecast the solar power. Finally, the two approaches are combined by normalizing the forecast with the clear sky model, and finally using this as input to the linear forecasting model, such that an ARX model is formed.

The paper is organized as follows. First the data and how it is preprocessed is described. The next section contains an outline of the clear sky model, followed by a section where all the forecasting models are described. Then an evaluation is given and the results are presented, followed by a discussion of the results and ideas for further work. Finally, the paper ends with a conclusion.

C.2 Data

The data used in this study consist of hourly mean values of solar power from a 4 kW-peak PV-system and NWP of global irradiance. The NWP are provided by the Danish Meteorological Institute using the HIRLAM mesoscale NWP model. The data covers the entire year 2006.

The time series of hourly observed solar power is

$$\{P_t; t = 1, \dots, N\} \quad (\text{C.1})$$

where $N = 8760$. The NWP have a calculation time of 4 hours, which is taken into consideration, such that e.g. the forecast from 2009-01-01 00:00 are only available from 2009-01-01 04:00. The NWP are provided in a time resolution of 3 hours. They are pre-processed into time series of hourly values, such that the most recent available forecast k hours ahead is selected each hour. The time series for a given k of the direct radiation is

$$\{G_{t+k|t}^{\text{NWP}}; t = 1, \dots, N\} \quad (\text{C.2})$$

C.2.1 Pre-processing

The solar power data is plotted for each hour of the day in Figure C.1. The solar radiation is zero at night, hence the observed solar power is also zero. For the current data set only periods, for a given hour of the day longer than 40 days in which the solar power is different from zero, are included for evaluation of the model performance. This is illustrated in Figure C.1, where the non-included periods are grayed out.

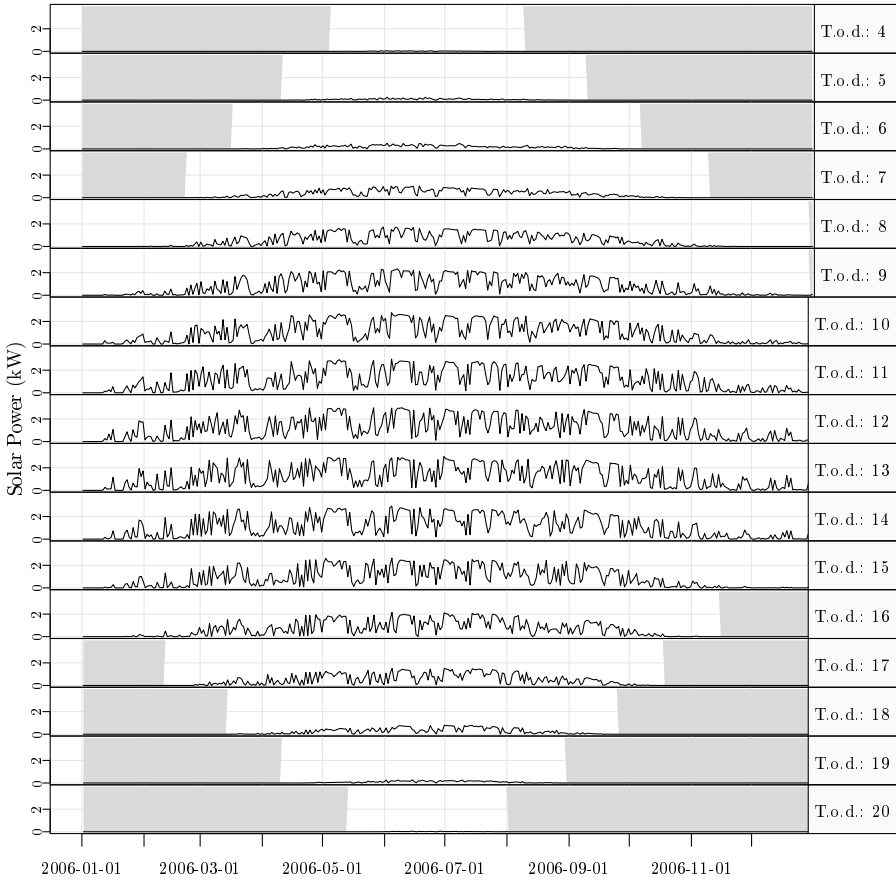


Figure C.1: The solar power data. The greyed area are not included in the evaluation of the model performance.

C.3 Clear sky model

Forecasting effectively using linear time series methods calls for stationarity of the underlying process Madsen (2007). The process that generates the solar power is not stationary, which is seen by plotting quantiles of the distribution of solar power conditioned on the time of day, see Figure C.2. Clearly the distribution of solar power is not independent of the time of day.

Most of this dependency can be removed by a normalization using a clear sky

model

$$\tau_t = \frac{P_t}{P_t^{\text{cs}}} \quad (\text{C.3})$$

where P_t is the observed solar power, P_t^{cs} is the estimated clear sky solar power, and τ_t is the normalized solar power.

C.3.1 Statistically estimated clear sky solar power

The clear sky solar power is estimated using a statistical non-linear and adaptive model. Quantile regression Koenker (2005) locally weighted in the day of year and time of day dimension is applied. This is carried out fully causal, i.e. only past values are used. The clear sky model is

$$\hat{P}_t^{\text{cs}} = q_{0.99}(P_1, P_2, \dots, P_t, h_y, h_{\text{tod}}) \quad (\text{C.4})$$

where $q_{0.99}$ is the 99% quantile based on the solar power values up to time t . The bandwidths h_{day} and h_{tod} , are in the day of year and time of day dimension, respectively. The bandwidths control how “locally” the model is fitted, i.e. a lower bandwidth puts more emphasis on data which is close in the two dimensions. The local weighting function is an Epanechnikov kernel. The applied bandwidths are

$$h_{\text{day}} = 100 \text{ days}, \quad h_{\text{tod}} = 3 \text{ hours} \quad (\text{C.5})$$

which were found by visual inspection of the fitted clear sky curve. Finally, it is noted that second-order polynomials were applied in the time of day dimension to include curvature into the model. The estimated clear sky solar power is shown in Figure C.3.

One advantage of the normalization is that it will automatically adapt to changes in the system, such as degraded performance or changes in the surroundings e.g. snow cover and shadowing effects. It can as well be used for monitoring of the solar system, since degraded performance from the same time of year will result in a lower clear sky solar power curve. Plots of the quantiles of the distribution of normalized solar power conditional on the time of day are shown in Figure C.2, from which it is seen that the normalized solar power process is considerably less dependent on the time of day and therefore a much more stationary process. It is noted that further work could include physical considerations into the clear sky model.

[t]

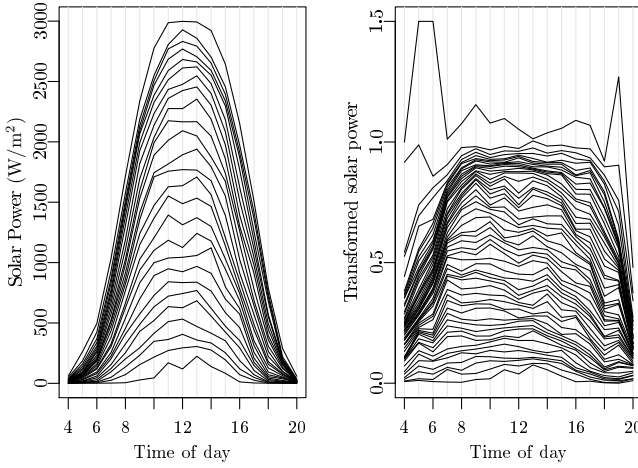


Figure C.2: The 0, 4%, ..., 100% quantiles of the distribution of the solar power and the normalized solar power conditioned on the time of day. Values above 1.5 has been clipped, which was the case for 6 values.

C.4 Forecasting models

In this section a description of the applied forecasting models is given. The models can be divided into models using linear time series models to forecast the normalized solar power: autoregressive (AR) and autoregressive with exogenous inputs (ARX) models - and models which forecast in a single stage: conditional parametric (CP) models. Each model is fitted separately for each horizon, such that the same model structure is used, but the parameters are estimated separately for each horizon.

C.4.1 Reference model

To compare the performance of prediction models, and especially when making comparisons between different studies, a common reference model is essential. The reference model for solar power used in this study is the best performing naive predictor for a given horizon. Two naive predictors of solar power are found to be relevant. Persistence

$$p_{t+k} = p_t + e_{t+k}, \quad (\text{C.6})$$

and diurnal persistence

$$p_{t+k} = p_{t-s(k)} + e_{t+k} \quad (\text{C.7})$$

$$s(k) = 24 + k \bmod 24 \quad (\text{C.8})$$

where $s(k)$ ensures that the latest diurnal observation is used, i.e. the value which, depending on the horizon, is either 24 or 48 hours before the time point that is to be forecasted.

C.4.2 Autoregressive models

Autoregressive (AR) models are applied to forecast the normalized solar power. These models can include either the latest available observation or the latest available diurnal observation, or both, as input. The models are fitted with k -step recursive least squares with forgetting factor Bacher et al. (2009). The model formulated as a k -step AR model

$$\tau_{t+k} = m + a_1 \tau_t + a_{24} \tau_{t-s(k)} + e_{t+k} \quad (\text{C.9})$$

$$s(k) = 24 + k \bmod 24 \quad (\text{C.10})$$

where the function $s(k)$ ensures that the latest observation of the diurnal component is included. The model without the diurnal component, denoted AR , performs best on short horizons

$$\tau_{t+k} = m + a_1 \tau_t + e_{t+k} \quad (\text{C.11})$$

and is included in the evaluation. The AR model with only the diurnal performs better on longer horizons, but is inferior to the models including the NWP.

C.4.3 Conditional parametric models

Conditional parametric (CP) models where the coefficients are conditional on the time of day and time of year are applied with both past solar power observations and NWPs as inputs. The CP model with the latest solar power observation as input is

$$P_{t+k} = m + a(t_{\text{day}}, t_{\text{tod}}, P_t) P_t + e_{t+k} \quad (\text{C.12})$$

where the coefficient function is a non-linear function of the solar power. It is denoted as CP_P . The CP model with NWPs of global radiation as input is

$$P_{t+k} = m + b(t_{\text{day}}, t_{\text{tod}}, G_{t+k|t}^{\text{nwp}}) G_{t+k|t}^{\text{nwp}} + e_{t+k} \quad (\text{C.13})$$

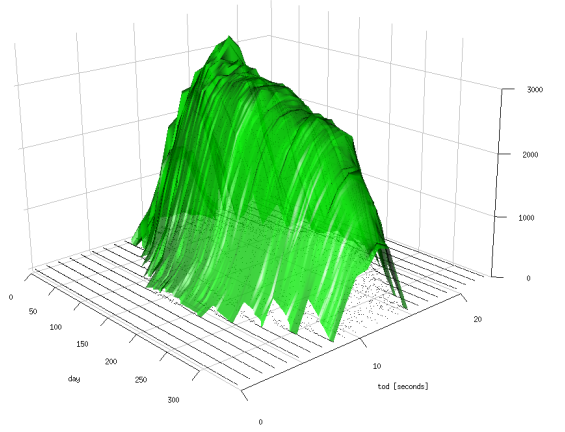


Figure C.3: The estimated clear sky solar power.

where $G_{t+k|t}^{\text{nwp}}$ is the k -hour ahead NWP of global radiation. This model is denoted CP_{NWP} . Finally, the model with both inputs

$$P_{t+k} = m + a(t_{\text{day}}, t_{\text{tod}}, P_t)P_t + b(t_{\text{day}}, t_{\text{tod}}, G_{t+k|t}^{\text{nwp}})G_{t+k|t}^{\text{nwp}} + e_{t+k} \quad (\text{C.14})$$

is denoted $CP_{NWP,P}$.

In the following the coefficients dependency of the time of day for CP_{NWP} is elaborated on. It is noted that the bandwidths are optimized for each horizon. Plots of the fitted forecasting function $b(t_{\text{day}}, t_{\text{tod}}, G_{t+k|t}^{\text{nwp}})$ for $k = 24$ hours are shown in Figure C.4. It is seen how the slope of the function is lower in the morning, than in the middle of the day. This is naturally caused by the higher angle of incidence in the morning, which cause less horizontal radiation to be absorbed due to reflection. Likewise for the afternoon. Finally, non-linearity in the fitted function is seen.

C.4.4 Autoregressive model with exogenous input

The AR model is be expanded to include the forecast of the CP models, thus combining information in past observed solar power and NWPs. The solar power

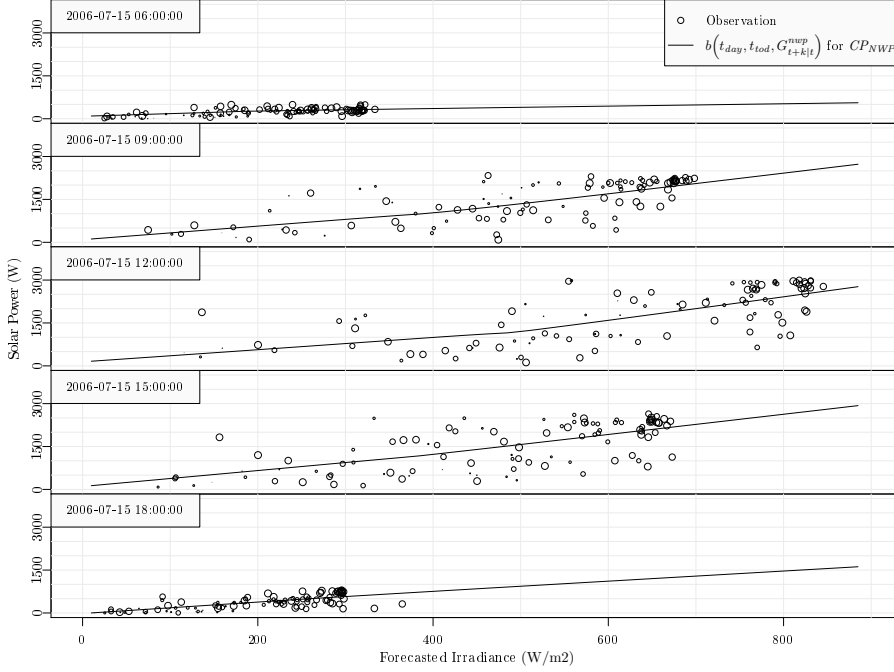


Figure C.4: Examples of the function fitted for $k = 24$ hours forecasting with the NWP's of global radiation at different times of the day on the 15'th of July 2010 with the CP_{NWP} model. For each observation the size of circle indicates the weighting of the observation in the CP models. Thus observations with a larger circle have more influence on the fitted function.

forecasts from the CP is normalized with the clear sky model by

$$\hat{\tau}_{t+k|t}^{\text{nwp}} = \frac{\hat{P}_{t+k|t}^{\text{nwp}}}{P_{t-s(k)}^{\text{cs}}} \quad (\text{C.15})$$

$$s(k) = f_{\text{spd}} + k \bmod f_{\text{spd}} \quad (\text{C.16})$$

where $f_{\text{spd}} = 24$ is the sample frequency in number of samples per day. The *ARX1* model is

$$\tau_{t+k} = m + a_1 \tau_t + b_1 \tau_{t+k|t}^{\text{nwp}} + e_{t+k} \quad (\text{C.17})$$

C.5 Evaluation

The methods used for evaluating the prediction models are inspired by Madsen et al. (2005). The clear sky model, RLS, and CP fitting do not use any degrees of freedom and the data set is therefore not divided into a training set and a test set. It is only for the optimization of the kernel bandwidths and the forgetting factor that the entire data set is used. The period before 2006-03-01 is considered as a burn-in period and not used for calculating the error measures.

C.5.1 Error measures

The Root Mean Square Error for the k 'th horizon is

$$RMSE_k = \left(\frac{1}{N} \sum_{t=1}^N e_{t+k}^2 \right)^{\frac{1}{2}} \quad (\text{C.18})$$

where e_{t+k} is the k -hourly prediction error. The $RMSE_k$ is used as the main evaluation criterion (EC) for the performance of the models. The Normalized Root Mean Square Error is found by

$$NRMSE_k = \frac{RMSE_k}{p_{\max}} \quad (\text{C.19})$$

where p_{\max} is the maximum observed solar power output. The mean value of the $RMSE_k$ for a range of horizons

$$\overline{RMSE}_{k_{\text{start}}, k_{\text{end}}} = \frac{1}{k_{\text{end}} - k_{\text{start}} + 1} \sum_{k=k_{\text{start}}}^{k_{\text{end}}} RMSE_k \quad (\text{C.20})$$

is used as a summary error measure. When comparing the performance of two models the improvement

$$I_{\text{EC}} = 100 \cdot \frac{EC_{\text{ref}} - EC}{EC_{\text{ref}}} (\%) \quad (\text{C.21})$$

is used, where EC is the considered evaluation criterion. When calculating the error measures it is important to consider how to handle missing values for the solar power forecasts. The problem is handled by replacing missing forecast values with forecast values from the reference model *Ref*.

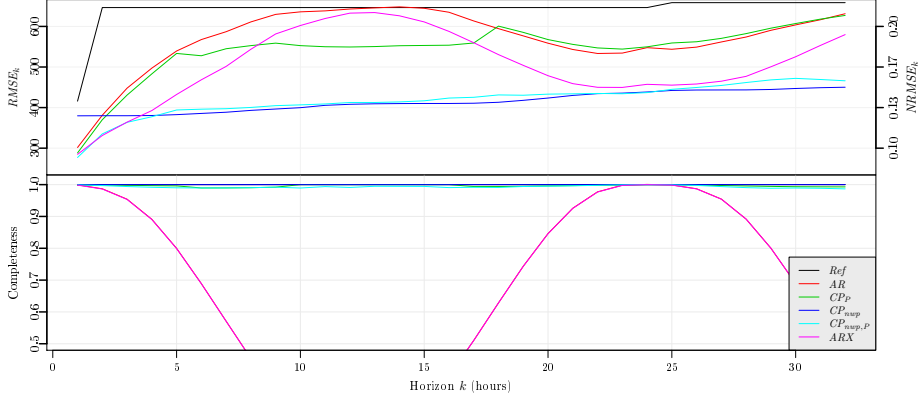


Figure C.5: The upper plot is $RMSE_k$ for the forecasting models. On the right side the $NRMSE_k$ is indicated. The lower plot is completeness C_k .

C.5.2 Completeness

In order to evaluate a model for its performance regarding missing forecast values a measure is defined. It is denoted completeness. The completeness of a forecast for horizon k , is the ratio of the total sum of solar power and the summed solar power for time points where the forecasts are not missing

$$C_k = \frac{\sum_{t=1}^N P_t I(\hat{P}_{t|t-k})}{\sum_{t=1}^N P_t} \quad (C.22)$$

where $I(\hat{P}_{t|t-k})$ is the indicator function which is 0 if $\hat{P}_{t|t-k}$ is missing, and 1 if not. Only the included values are used, i.e. not values during nighttime.

C.6 Results

In this section the results are presented and evaluated. The $\overline{RMSE}_{k_{start}, k_{end}}$ improvements for relevant ranges of horizons are listed in Table C.1. For selected models the $RMSE_k$ is shown in the upper plot of Figure C.5 and the completeness in the lower.

Considering the improvements it is seen that most of the models perform very well on either the short horizons or the longer horizons. Starting with short horizons (1 to 2 hours) the four models using the latest observed solar power

have better performance than CP_{NWP} , which only uses the NWPs. Using the combination of observed solar power and NWPs improves the performance, except on longer horizons where using only NWPs are slightly better. Considering the performance of AR , CP_P , and ARX it is seen that the $RMSE_k$ increase really fast as the horizon increases and reach the reference model around a horizon of 10 hours. This is simply because the models are using night values (which are missing) to forecast day values. This is also seen in the completeness of the AR and ARX model.

C.7 Discussion and applications

This section contains a short discussion of the results and ideas for further work, and ends with an outline of applications.

Considering the improvement achieved over the reference model the forecasting models are found to perform very well. Clearly the quality of the NWPs of solar radiation is the most influential source of error, hence improved NWPs will improve the performance. Especially using NWPs of direct and diffuse radiation should be tried. Regarding further improvement of the forecasting models, it is suggested that the following should be considered:

- Application of regime models and hidden Markov models to handle different aspects of forecasting for e.g. low and high radiation values, and it might be useful to use different forecasting models for different types of cloud conditions. This is ideal to apply in the setting of the CP models.
- For the CP models using higher order polynomials in the day of year and time of day dimensions should improve the models. It was tried but didn't improve the performance, but as the NWPs are getting better this will most likely be important.

Table C.1: Improvements in percent for selected ranges of horizons.

Model	$I_{RMSE_{1,2}}$	$I_{RMSE_{9,17}}$	$I_{RMSE_{18,32}}$
AR	34.3	7.4	12.6
CP_P	36.7	17	11.5
CP_{NWP}	25	38.4	33.1
$CP_{NWP,P}$	40.8	37.6	31.4
ARX	40.1	15.9	25

- A thorough evaluation of the forecast errors to find ideas for how the models can be improved.

The applications for solar power forecasting include the integration of PV systems into the electricity grid, especially for smart grids. The solar power forecasts can be used as input to model predictive control to optimize the operation of the PV system. This will enable diurnal peak-shaving and cost optimization for smaller systems with storage capacity in battery packs (e.g. provided in an electrical car). For large solar power producers forecasting is essential for optimized energy trading.

The method is furthermore well suited for monitoring the performance of PV systems. Measures of the performance can be derived from the CP models, with which systems can be compared on an absolute scale. Sudden high deviation from the CP forecasting model will allow for very fast detection of failures in the system. For an individual system the change in performance over time can also be assessed by monitoring the clear sky curve for unusual behavior, and compare the change from year to year.

C.8 Conclusion

Two approaches for solar power forecasting are presented and applied to forecast hourly values for horizons up to 32 hours. Both a method based on a two-stage approach, where first the solar power is normalized with a statistical clear-sky model, and a method in which the solar power is forecasted in a single step. The normalization with a clear sky model removes most of the non-stationarity caused by the changing position of the sun relative to the PV panel. This is a pre-requisite for optimal application of linear time series models. Conditional parametric models are used to include NWP's of global radiation, and a one-stage approach, solely based on conditional parametric models, is presented. A root mean square improvement over a persistence reference model on short horizons (1 to 2 hours) is in average 40%, and in average 35% on the longer horizons. The method can furthermore be applied to monitor and check the performance of PV systems.

References

- P. Bacher, H. Madsen, and H. A. Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83(10):1772–1783, 2009. ISSN 0038092x.

- F. O. Hocaoglu, Ö. N. Gerek, and M. Kurban. Hourly solar radiation forecasting using optimal coefficient 2-d linear filters and feed-forward neural networks. *Solar Energy*, 82(8):714–726, 2008. ISSN 0038092x.
- W. Ji and K. C. Chee. Prediction of hourly solar radiation using a novel hybrid model of arma and tdnn. *Solar Energy*, 85(5):808–817, 2011. ISSN 0038092x. doi: 10.1016/j.solener.2011.01.013.
- R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.
- E. Lorenz, J. Hurka, D. Heinemann, and H. Beyer. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2(1): 2–10, 2009. ISSN 19391404, 21511535.
- H. Madsen. *Time Series Analysis*. Chapman & Hall, 2007.
- H. Madsen, P. Pinson, G. Kariniotakis, H. A. Nielsen, and T. S. Nielsen. Standardizing the performance evaluation of shortterm wind power prediction models. *Wind Engineering*, 29(6):475, 2005. ISSN 0309524x.
- T. S. Nielsen, H. Madsen, and H. A. Nielsen. Prediction of wind power using time-varying coefficient-functions. *Proceedings of the 15th IFAC World Congress on Automatic Control*, 2002.
- A. Sfetsos and A. Coonick. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. *Solar Energy*, 68(2): 169–178, 2000. ISSN 0038092x.

P A P E R D

Short-term solar collector power forecasting

Authors:

Peder Bacher¹, Henrik Madsen¹, and Bengt Perers²

In proceedings:

ISES Solar World Congress 2011.

¹DTU Informatics, Richard Pedersens Plads, Building 321, DK-2800 Lyngby, Denmark

²DTU Civil Engineering, Brovej, Building 118, DK-2800 Lyngby, Denmark

Abstract

This paper describes a new approach to online forecasting of power output from solar thermal collectors. The method is suited for online forecasting in many applications and in this paper it is applied to predict hourly values of power from a standard single glazed large area flat plate collector. The method is applied for horizons of up to 42 hours.

Solar heating systems naturally come with a hot water tank, which can be utilized for energy storage also for other energy sources. Thereby such systems can become an important part of energy systems with a large share of uncontrollable energy sources, such as wind power. In such a scenario online forecasting is a vital tool for optimal control and utilization of solar heating systems.

The method is a two-step scheme, where first a non-linear model is applied to transform the solar power into a stationary process, which then is forecasted with robust time-adaptive linear models. The approach is similar to the one by Bacher et al. (2009), but contains additional effects due to differences between solar thermal collectors and photovoltaics. Numerical weather predictions provided by Danish Meteorological Institute are used as input. The applied models adapt over time enabling tracking of changes in the system and in the surrounding conditions, such as decreasing performance due to wear and dirt, and seasonal changes such as leaves on trees. This furthermore facilitates remote monitoring and check of the system.

D.1 Introduction

Forecasting of energy production is vital for optimization of energy systems which include wind and solar energy production. This paper describes an approach to online forecasting of power production from solar thermal collectors. In Denmark the level of wind power penetration already now gives periods with a surplus of energy and facilities to absorb this energy are needed. Solar heating systems with a hot water tank and auxiliary electrical heating can provide energy storage, which can facilitate absorption of wind energy and peak shaving, especially for levelling out diurnal energy consumption. The method is planned to be part of the control system for such heating systems (Perers et al., 2011). The study is carried out with climate data observed at a weather station at Danish Technical University. From this data, simulated hourly average values of solar thermal power is generated with a very detailed simulation model. Furthermore numerical weather predictions (NWP) provided by Danish Meteorological In-

stitute data is used. The forecasting method is a two-step scheme, where first a statistical clear sky model is applied to transform the solar power into a more stationary process, which then is forecasted with robust time-adaptive linear models. The NWP's are used as input to conditional parametric time-adaptive models to forecast the solar power. These forecasts are then transformed with the clear sky model, such that they can be applied as inputs to the linear forecasting models. Finally, a combined model, which is the most optimal for all horizons, is formed.

The paper is organized as follows. First the data and how it is preprocessed is described in a section. The next section contains an outline of the clear sky model, and this is followed by a section where all the forecasting models are described. Then an evaluation is given and the results are presented, each in a section. The second last section contains a discussion of the results and ideas for further work, and finally, the paper ends with a conclusion.

Nomenclature

P_t	Hourly solar thermal power, $[\text{W}/\text{m}^2]$.
P_t^{cs}	Estimated clear sky solar power, $[\text{W}/\text{m}^2]$.
τ_t	Normalized solar power.
t	Time index, [h].
k	Forecast horizon index, [h].
t_{tod}	Time of day.
$F'(\tau\alpha)_{\text{en}}$	Zero loss efficiency of collector for direct radiation at normal incidence
$K_{\tau\alpha\text{b}}(\theta)$	Incidence angle modifier for direct radiation
$K_{\tau\alpha\text{d}}$	Incidence angle modifier for diffuse radiation
$F'U_0$	Heat loss coefficient at $(T_{\text{a}} - T_{\text{f}}) = 0$, $[\text{W}/(\text{m}^2\text{K})]$.
$G_{t+k t}^{\text{nwp}}$	NWP of global radiation, $[\text{W}/\text{m}^2]$.
$G_{t+k t}^{\text{b,nwp}}$	NWP of direct solar radiation, $[\text{W}/\text{m}^2]$.
$G_{t+k t}^{\text{d,nwp}}$	NWP of diffuse solar radiation, $[\text{W}/\text{m}^2]$.
$T_{t+k t}^{\text{a,nwp}}$	NWP of ambient temperature, $[\text{°C}]$.
$\hat{P}_{t+k t}^{\text{nwp}}$	k -hour prediction of solar power, $[\text{W}/\text{m}^2]$.
$\hat{\tau}_{t+k t}$	k -hour prediction of normalized solar power.
e_{t+k}	k -step prediction error.

D.2 Data

The forecasting method is applied on simulated solar output power data for a flat plate collector carefully tested and modelled at DTU. A validated collector model and longterm climate data from the DTU Byg climate station was used to create realistic operating data for a solar collector during the year. The simulation model and weather data was introduced in TRNSYS 16 and the collector output power was calculated as hourly mean values. The simulation model is dynamic, such that dynamical effects - introduced when the collector starts and stops and during rapidly varying solar radiation conditions - are modelled.

In this study all time series are hourly average values and all units are implicitly per hour. Time points are set to the end of their respective sample period and all are in UTC. The units for radiation are $\frac{\text{W}}{\text{m}^2}$ and for temperatures $^{\circ}\text{C}$.

D.2.1 Solar power

The simulated solar power time series is plotted for each hour of the day in Figure D.1. A few short periods are with missing values.

D.2.2 Numerical weather predictions

The numerical weather predictions (NWP) used in the study are provided by Danish Meteorological Institute. The NWP model is DMI-HIRLAM-S05, which has a 5 kilometer grid and 40 vertical layers (DMI, 2011). NWP are updated every 6'th hour and are up to a 48 hours horizon. They consist of hourly predictions of ambient temperature, and horizontal direct- and diffuse solar radiation. A couple of the considered forecasting models use the global radiation as input, which is simply the direct and the diffuse radiation summed. The scatter plots in Figure D.2 shows the solar power versus the NWP for a 24 hour horizon. Clearly, the solar power is highly correlated with both the global and direct radiation, whereas the effect of diffuse and ambient temperature are not as apparent.

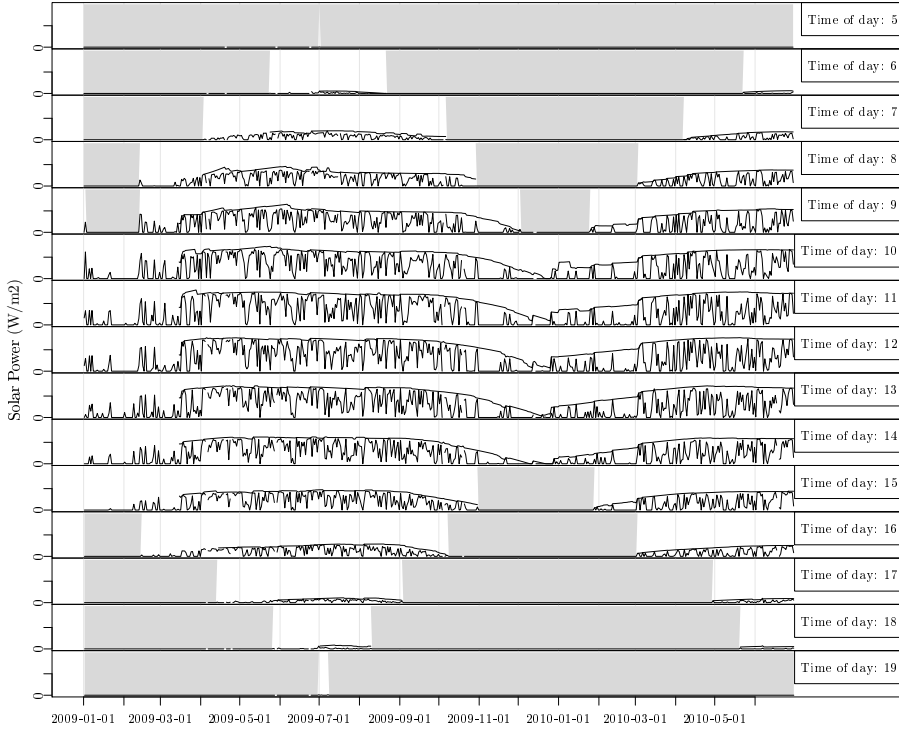


Figure D.1: The observed solar power for each hour of the day. The grayed area shows the periods not included in the modelling. The curve following the highest values of solar power is the estimated clear sky power, which is referred to in later parts of the paper.

D.2.3 Pre-processing

On most locations on earth the solar radiation is zero at night time, hence the observed solar power is also zero. For the current dataset only periods, for a given hour of the day longer than 40 days in which the solar power is different from zero, are included. This is illustrated in Figure D.1, where the non-included periods are grayed out. Furthermore a few short periods are missing from the observations. The time series of hourly observed solar power spanning the period from 2009-01-01 to 2010-07-01 is

$$\{P_t; t = 1, \dots, N\} \quad (\text{D.1})$$

where $N = 13104$. The NWP's have a calculation time of 4 hours, which is taken into consideration, such that e.g. the forecast from 2009-01-01 00:00 are only available from 2009-01-01 04:00. The NWP's are pre-processed into time series

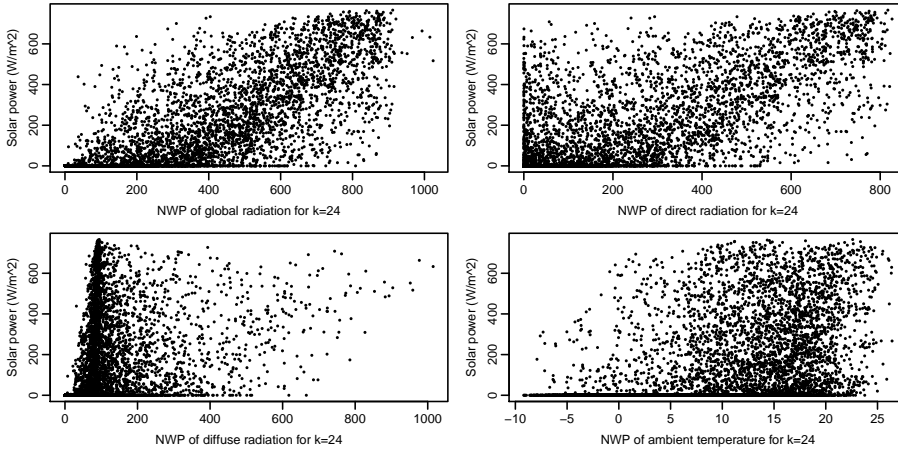


Figure D.2: The observed solar power versus the NWPs for $k = 24$, which are used as inputs to forecasting models.

of hourly values, such that the most recent available forecast k hours ahead is selected each hour. The time series are for a given k : the direct radiation

$$\left\{ G_{t+k|t}^{\text{b,nwp}}; t = 1, \dots, N \right\} \quad (\text{D.2})$$

the diffuse radiation

$$\left\{ G_{t+k|t}^{\text{d,nwp}}; t = 1, \dots, N \right\} \quad (\text{D.3})$$

and the ambient temperature

$$\left\{ T_{t+k|t}^{\text{a,nwp}}; t = 1, \dots, N \right\} \quad (\text{D.4})$$

Due to the 6 hours interval the NWPs for horizons longer than 42 hours are not complete and therefore the solar power forecasting are only carried out up to 42 hours.

D.3 Clear sky model

For effective forecasting with classical linear time series methods stationarity of the process is required (Madsen, 2007). The process that generates the solar power is not stationary, which is seen by plotting quantiles of the distribution of solar power conditioned on the time of day. Such a plot is shown in Figure D.3.

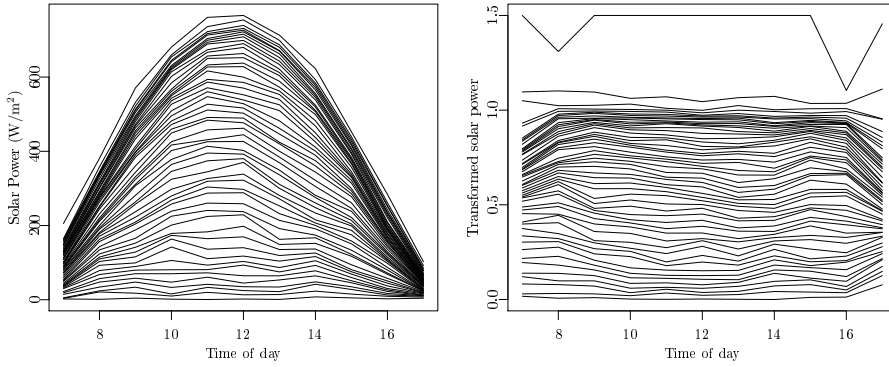


Figure D.3: The 0, 2%, ..., 100% quantiles of the distribution of the solar power and the transformed solar power conditioned on the time of day. Values above 1.5 has been clipped, which was the case for 17 values.

Clearly the distribution of solar power is not independent of the time of day. The dependency can be removed by a transformation with a clear sky model

$$\tau_t = \frac{P_t}{P_t^{\text{cs}}} \quad (\text{D.5})$$

where P_t is the observed solar power, P_t^{cs} is the estimated clear sky solar power, and τ_t is the transformed solar power.

D.3.1 Statistically estimated clear sky solar power

The clear sky solar power is estimated with a statistical non-linear and adaptive model. Quantile regression locally weighted in the day of year and time of day dimension is applied. In the present study this is carried out fully causal. The clear sky model is

$$P_t^{\text{cs}} = q_{0.99}(P_1, P_2, \dots, P_t, h_t, h_{\text{tod}}, h_y) \quad (\text{D.6})$$

where $q_{0.99}$ is the 99% quantile of all the solar power values up to t . The bandwidths h_t , h_{tod} , and h_y , are in the time-, time of day-, and year-dimension, respectively. The bandwidths control how “locally” the model is fitted, i.e. a lower bandwidth puts more emphasis on data which is close in time. The local weighting function is an Epanechnikov kernel. The applied bandwidths are

$$h_t = 120 \text{ days}, \quad h_{\text{tod}} = 2 \text{ hours}, \quad h_y = 1.7 \text{ years} \quad (\text{D.7})$$

which were found by visual inspection of the fitted clear sky curve. Finally, it is noted that second-order polynomials were applied in the time- and time of day-dimension to include curvature into the model. The estimate of the clear sky solar power is shown in Figure D.1. From the plot it is seen that it follows the highest values of solar power quite well. Clearly, the clear sky power is most easily carried out in the periods with a high level of solar power. One advantage of the transformation is that it will automatically adapt to changes in the system, such as degraded performance or changes in the surroundings e.g. snow cover and shadowing effects. It can as well be used for monitoring of the solar system, since degraded performance from the same time of year will result in a lower clear sky solar power curve. Quantile plots of the transformed solar power conditioned on the time of day are shown in Figure D.3, from which it is seen that the transformed solar power process is considerably less dependent of the time of day and therefore a much more stationary process. It is noted that further work could include physical considerations e.g. by using the air mass as an input.

D.4 Forecasting models

In this section a description of the applied forecasting models is given. The models can be divided into models without NWPs as input - autoregressive (AR) models - and models with NWPs as input: conditional parametric (CP) and autoregressive with exogenous inputs (ARX) models. Each model is fitted separately for each horizon, such that the same model structure is used, but the parameters are estimated separately for each horizon. In the final model, a combination of models are used to achieve the most optimal performance for all horizon.

D.4.1 Reference model

To compare the performance of prediction models, and especially when making comparisons between different studies, a common reference model is essential. The reference model for solar power used in this study is the best performing naive predictor for a given horizon. Two naive predictors of solar power are found to be relevant. Persistence

$$p_{t+k} = p_t + e_{t+k}, \quad (\text{D.8})$$

and diurnal persistence

$$p_{t+k} = p_{t-s(k)} + e_{t+k} \quad (\text{D.9})$$

$$s(k) = f_{\text{spd}} + k \bmod f_{\text{spd}} \quad (\text{D.10})$$

where $f_{\text{spd}} = 24$ is the sample frequency in number of samples per day and $s(k)$ ensures that the latest diurnal observation is used, i.e. the value which, depending on the horizon, is either 24 or 48 hours before the time point that is to be forecasted.

D.4.2 Autoregressive models

Autoregressive (AR) models are applied to forecast the transformed solar power. These models can include either the latest available observation or the latest available diurnal observation, or both, as input. The models are fitted with k -step recursive least squares with forgetting factor (Bacher et al., 2009). The model formulated as a k -step AR model

$$\tau_{t+k} = m + a_1 \tau_t + a_{24} \tau_{t-s(k)} + e_{t+k} \quad (\text{D.11})$$

$$s(k) = 24 + k \bmod 24 \quad (\text{D.12})$$

where the function $s(k)$ ensures that the latest observation of the diurnal component is included. It was found that depending on the horizon better performance was achieved by only using one input. Thus for short horizons (1 and 2 hours) the model without the diurnal component

$$\tau_{t+k} = m + a_1 \tau_t + e_{t+k} \quad (\text{D.13})$$

was found to have the best performance, it is denoted *AR1*, and for longer horizons the model with only the diurnal component

$$\tau_{t+k} = m + a_{24} \tau_{t-s(k)} + e_{t+k} \quad (\text{D.14})$$

was found to have the best performance, it is denoted *ARDiurnal*.

D.4.3 Conditional parametric models with NWP as input

Models based on NWPs of solar radiation and ambient temperature are described in this section. It is known from physics (Perers, 1997) that the power

output of a solar collector can be described by

$$P = F'(\tau\alpha)_{\text{en}} K_{\tau\alpha b}(\theta) G_{\text{b,col}} + F'(\tau\alpha)_{\text{en}} K_{\tau\alpha d} G_{\text{d,col}} - F'U_0 \left(\frac{T_o + T_i}{2} - T_a \right) \quad (\text{D.15})$$

where the $G_{\text{b,col}}$ and $G_{\text{d,col}}$ are respectively direct and diffuse solar radiation normal to the collector plane. This is formed into a forecasting model based on NWP by rewriting as follows. First, both the angle of incidence modifier $K_{\tau\alpha b}(\theta)$ and the transformation of solar radiation from horizontal to the collector plane are modelled by letting the coefficients - for the radiation effects - become a function of time t and time of day t_{tod} . Furthermore, assuming that the outlet temperature is a function of the solar radiation

$$T_o = f_b(G_{\text{b,col}}) + f_d(G_{\text{d,col}}) \quad (\text{D.16})$$

this give the total effect of direct radiation as a non-linear function

$$a(t, t_{\text{tod}}, G_{\text{b}}, G_{\text{d}}) G_{\text{b}} = F'(\tau\alpha)_{\text{en}} K_{\tau\alpha b}(\theta) G_{\text{b,col}} - F'U_0 \frac{1}{2} f_b(G_{\text{b,col}}) \quad (\text{D.17})$$

and for the diffuse radiation

$$b(t, t_{\text{tod}}, G_{\text{b}}, G_{\text{d}}) G_{\text{d}} = F'(\tau\alpha)_{\text{en}} K_{\tau\alpha d} G_{\text{b,col}} - F'U_0 \frac{1}{2} f_d(G_{\text{d,col}}) \quad (\text{D.18})$$

Finally, the effect of the ambient temperature is kept as

$$cT_a = F'U_0 T_a \quad (\text{D.19})$$

and by assuming a constant inlet temperature this part becomes a constant effect

$$m = -F'U_0 \frac{T_i}{2} \quad (\text{D.20})$$

Thus the CP model structure used for forecasting is

$$P = m + a(t, t_{\text{tod}}, G_{\text{b}}, G_{\text{d}}) G_{\text{b}} + b(t, t_{\text{tod}}, G_{\text{b}}, G_{\text{d}}) G_{\text{d}} + cT_a \quad (\text{D.21})$$

Since the time-dependency and non-linearity are smooth functions in the parameters, it is modelled with conditional parametric (CP) models. The time varying effect is modelled by conditioning on t and t_{tod} - this is equivalent of a local constant effect - and the dependency of the radiation is modelled with 1-order local polynomials. A kernel method is applied, using a nearest neighbor approach to find the bandwidth of an Epanechnikov weighting function. From Equation (D.15) it is seen that the output can be negative if little radiation hits the collector and the ambient temperature is low. In this case the output is zero

since the system stops. This effect can be seen on the plot in Figure D.2. It is handled by the non-linearity of the models and by setting negative forecasts to zero.

The simplest considered conditional parametric model is

$$P_{t+k} = m + a\left(t, t_{\text{tod}}, G_{t+k|t}^{\text{nwp}}\right) G_{t+k|t}^{\text{nwp}} + e_{t+k} \quad (\text{D.22})$$

where $G_{t+k|t}^{\text{nwp}}$ is the k -hour NWP of global radiation and denoted as *CP1* in the following. This second CP model has NWPs of direct and the diffuse radiation as inputs

$$\begin{aligned} P_{t+k} = m + a\left(t, t_{\text{tod}}, G_{t+k|t}^{\text{b,nwp}}, G_{t+k|t}^{\text{d,nwp}}\right) G_{t+k|t}^{\text{b,nwp}} \\ + b\left(t, t_{\text{tod}}, G_{t+k|t}^{\text{b,nwp}}, G_{t+k|t}^{\text{d,nwp}}\right) G_{t+k|t}^{\text{d,nwp}} + e_{t+k} \end{aligned} \quad (\text{D.23})$$

where $G_{t+k|t}^{\text{b,nwp}}$ is the k -hour NWP of direct radiation and $G_{t+k|t}^{\text{d,nwp}}$ is the k -hour NWP of diffuse radiation, and denoted as *CP2*. Finally the model is expanded with NWPs of ambient temperature

$$\begin{aligned} P_{t+k} = m + a\left(t, t_{\text{tod}}, G_{t+k|t}^{\text{b,nwp}}, G_{t+k|t}^{\text{d,nwp}}\right) G_{t+k|t}^{\text{b,nwp}} \\ + b\left(t, t_{\text{tod}}, G_{t+k|t}^{\text{b,nwp}}, G_{t+k|t}^{\text{d,nwp}}\right) G_{t+k|t}^{\text{d,nwp}} + c\left(t, t_{\text{tod}}\right) T_{t+k|t}^{\text{a,nwp}} + e_{t+k} \end{aligned} \quad (\text{D.24})$$

where $T_{t+k|t}^{\text{a,nwp}}$ is the k -hour NWPs of the ambient temperature and the model is denoted as *CP3*.

In the following the coefficients dependency of the time of day is elaborated on. Plots of the fitted forecasting function $a\left(t, t_{\text{tod}}, G_{t+k|t}^{\text{b,nwp}}, G_{t+k|t}^{\text{d,nwp}}\right)$ are shown in Figure D.4. It is seen how the slope of the function is lower in the morning, than in the middle of the day. This is naturally caused by the higher angle of incidence in the morning, which cause less horizontal radiation to be absorbed due to reflection. Likewise for the afternoon. Finally, non-linearity in the fit is seen, which is caused by the non-negativity of the solar power (mentioned above) and varying uncertainty of the NWPs.

D.4.4 Autoregressive model with exogenous input

The AR model is be expanded to include the forecast of the CP models, thus combining information in past observed solar power and NWPs. The solar power

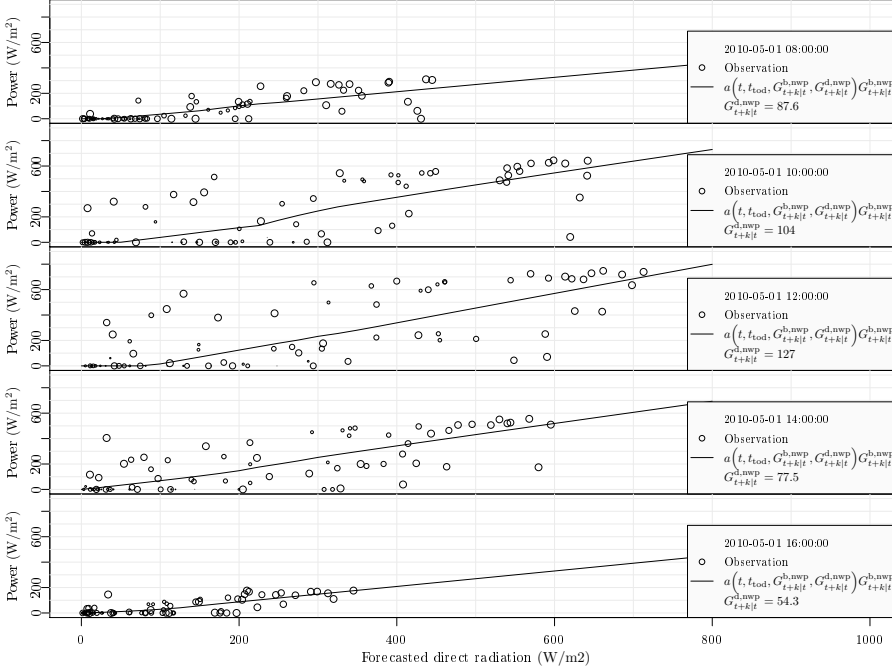


Figure D.4: Examples of the function fitted for forecasting of the effect of direct radiation at different times of the day the 1st of May 2010 with the *CP2* model. For each observation the size of circle indicates the weighting of the observation in the CP models. Thus observations with a larger circle have more influence on the fitted function.

forecasts from the CP is transformed with the clear sky model by

$$\hat{\tau}_{t+k|t}^{nwp} = \frac{\hat{P}_{t+k|t}^{nwp}}{P_{t-s(k)}^{cs}} \quad (D.25)$$

$$s(k) = f_{spd} + k \bmod f_{spd} \quad (D.26)$$

where $f_{spd} = 24$ is the sample frequency in number of samples per day. This is applied as an input to the ARX model

$$\tau_{t+k} = m + a_1 \tau_t + a_{24} \tau_{t-s(k)} + b_1 \tau_{t+k|t}^{nwp} + e_{t+k} \quad (D.27)$$

Again, as for the AR models, different performance is found depending on the horizon. The *ARX1* model is best for short horizons

$$\tau_{t+k} = m + a_1 \tau_t + b_1 \tau_{t+k|t}^{nwp} + e_{t+k} \quad (D.28)$$

ARXDiurnal for horizons up to 24 hours

$$\tau_{t+k} = m + a_{24}\tau_{t-s(k)} + b_1\tau_{t+k|t}^{\text{nwp}} + e_{t+k} \quad (\text{D.29})$$

and *ARX*

$$\tau_{t+k} = m + b_1\tau_{t+k|t}^{\text{nwp}} + e_{t+k} \quad (\text{D.30})$$

for longer horizons.

D.4.5 Combined model

The final model is a combination of the previously described models. The model is denoted *ARXCombined*. First, missing values in forecasts from *ARX1* are replaced with forecast values from *ARXDiurnal*. These missing values are in the morning, since they were tried to be forecasted based on night values, which are zero. For horizons longer than 30 hours forecasts from *ARX* are used. Finally, any remaining missing values - which are only where the diurnal lag was not present for *ARXDiurnal* - are replaced with forecasted values from *CP2*.

D.5 Evaluation

The methods used for evaluating the prediction models are inspired by Madsen et al. (2005). They suggest a framework for evaluation of wind power forecasting. The clear sky model, RLS, and CP fitting do not use any degrees of freedom and the dataset is therefore not divided into a training set and a test set. It is only for the optimization of the kernel bandwidths and the forgetting factor that the entire dataset is used. The period before 2009-03-15 is considered as a burn-in period and are not used when calculating the error measures.

D.5.1 Error measures

The Root Mean Square Error for the k 'th horizon is

$$RMSE_k = \left(\frac{1}{N} \sum_{t=1}^N e_{t+k}^2 \right)^{\frac{1}{2}} \quad (\text{D.31})$$

where e_{t+k} is k -hourly prediction error. The $RMSE_k$ is used as the main evaluation criterion (EC) for the performance of the models. The Normalized Root Mean Square Error is found by

$$NRMSE_k = \frac{RMSE_k}{p_{\max}} \quad (D.32)$$

where p_{\max} is the maximum observed solar power output. The mean value of the $RMSE_k$ for a range of horizons

$$\overline{RMSE}_{k_{\text{start}}, k_{\text{end}}} = \frac{1}{k_{\text{end}} - k_{\text{start}} + 1} \sum_{k=k_{\text{start}}}^{k_{\text{end}}} RMSE_k \quad (D.33)$$

is used as a summary error measure. When comparing the performance of two models the improvement

$$I_{\text{EC}} = 100 \cdot \frac{EC_{\text{ref}} - EC}{EC_{\text{ref}}} (\%) \quad (D.34)$$

is used, where EC is the considered evaluation criterion. When calculating the error measures it is important to consider how to handle missing values for the solar power forecasts. The problem is handled by replacing missing forecast values with forecast values from the reference model *Ref*.

D.5.2 Completeness

In order to evaluate a model for its performance regarding missing forecast values a measure is defined, it is denoted completeness. The completeness of a forecast for horizon k , is the ratio of the the summed solar power for time points where the forecasts are not missing to the total sum of solar power

$$C_k = \frac{\sum_{t=1}^N P_t I(\hat{P}_{t|t-k})}{\sum_{t=1}^N P_t} \quad (D.35)$$

where $I(\hat{P}_{t|t-k})$ is the indicator function which is 0 if $\hat{P}_{t|t-k}$ is missing, and 1 if not. Only the included values are used, i.e. not night values.

D.6 Results

In this section the results are presented and evaluated. The $\overline{RMSE}_{k_{\text{start}}, k_{\text{end}}}$ improvement for relevant ranges of horizons are listed in Table D.1. For selected models the $RMSE_k$ is shown in the upper plot of Figure D.5 and the completeness in the lower.

Considering the improvements it is seen that most of the models perform very well on either the short horizons or the longer horizons. Starting with short horizons (1 to 2 hours), the *AR1* and *ARX1* are clearly superior, which is due to their inclusion of the most present autoregressive lag. Their performance on longer horizons are not good. The reason for this is found by considering the plot of $RMSE_k$ and completeness. Here it is seen that the completeness of *AR1* and *ARX1* drops really quickly as the horizon increase, which cause the $RMSE_k$ to increase and reach the reference model at the 10 hours horizon. This is simply due to missing forecast values, since for e.g. the 10 to 14 hours horizons the models use night values (which are missing) to forecast day values with.

For horizons longer than three hours the best performance is seen for the models, that doesn't include the most present AR lag. The *ARDiurnal* is a clear improvement from the *AR1*, and the CP and ARX models - which include the NWP's - are superior for these horizons. An improved performance is found from *CP1* to *CP2* mainly for 3 to 24 hours horizons, whereas no clear increase in improvement is found from *CP2* to *CP3*. The CP models are slightly improved by using them as input to the ARX models, since autocorrelation of the errors are modelled.

Finally, the combined model *ARXCombined* utilizes the best parts of: *ARX1*, *ARX*, *ARXDiurnal*, and *CP2*. Especially the replacement of missing forecast values improves the performance for horizons up to 5 hours. The completeness of the combined model is as high as any of the others.

Table D.1: Improvements in percent for selected ranges of horizons.

Model	$I_{RMSE_{1,2}}$	$I_{RMSE_{3,24}}$	$I_{RMSE_{25,42}}$	$I_{RMSE_{1,42}}$
<i>AR1</i>	30.8	7.1	6.1	7.8
<i>ARDiurnal</i>	-10.7	15	18	15.1
<i>CP1</i>	13.5	30	30.6	29.5
<i>CP2</i>	16.2	31.4	30.9	30.5
<i>CP3</i>	15.5	31.6	30.8	30.5
<i>ARX</i>	17.1	32	31.3	31
<i>ARX1</i>	34.4	11.6	8.7	11.4
<i>ARXDiurnal</i>	17.8	32.4	30.5	30.9
<i>ARXCombined</i>	39.3	33.3	31.5	32.8

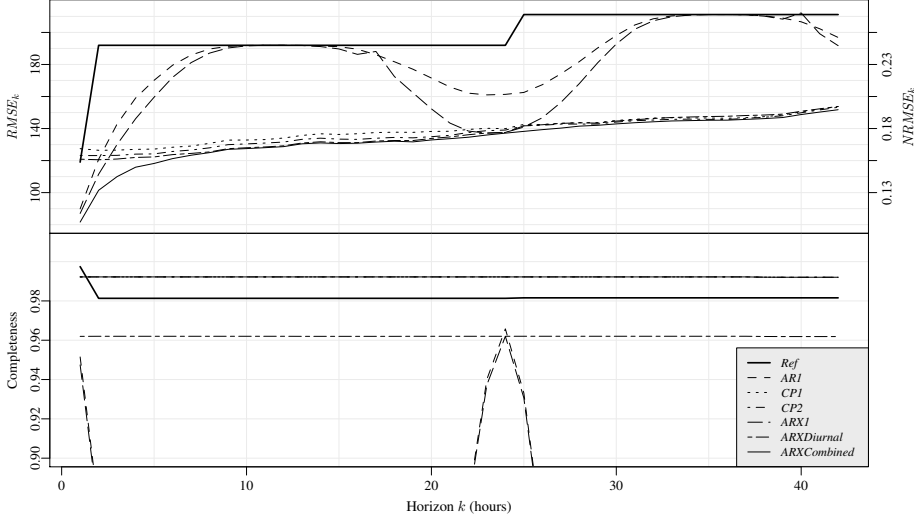


Figure D.5: The upper plot is $RMSE_k$ for the forecasting models. On the right side the $NRMSE_k$ is indicated. The lower plot is completeness C_k .

D.7 Discussion and applications

This section contains a short discussion of the results and ideas for further work, and ends with an outline of applications.

Considering the improvement achieved over the reference model the forecasting method is found to perform very well. Clearly the quality of the NWP of solar radiation is the most influential source of error, hence improved NWPs will improve the forecasting performance. Regarding improvement of the forecasting models, the following are considered:

- A thorough evaluation of the forecast errors to clarify how the models can be improved.
- Optimization of forgetting factor of the RLS has not been carried out, this will improve the performance of the AR and ARX models.
- Application of regime models to handle different aspects of forecasting for low and high radiation values.
- More optimal ways to combine the models. Utilizing a linear combination instead of the simple replacement currently carried out.

- It might be possible to improve performance by including a third-stage, where modelling of the errors are carried out.

The applications for this type of solar thermal power forecasting counts the integration of solar thermal energy systems with auxiliary electrical heating into smart grid systems (Perers et al., 2011). The solar power forecasts will be used for model predictive control to optimize the operation of the system. Other applications include optimal control of large solar heating plants.

The method is furthermore well suited for monitoring the performance of solar thermal systems. Measures of the performance can be derived from the CP models, with which systems can be compared on an absolute scale. Sudden high deviation from the CP forecasting model will allow for very fast detection of failures in the system. For an individual system the change in performance over time can also be assessed by monitoring the clear sky curve for unusual behavior, and compare the change from year to year.

D.8 Conclusion

A method for forecasting of solar thermal power output is presented. It is applied to forecast hourly values for horizons up to 42 hours. The method is based on conditional parametric models. Both models without and with NWP of solar radiation and ambient temperature are considered. The NWPs are included by using a non-linear conditional parametric model, which are formed from prior physical knowledge. The forecast models which do not use NWPs achieve an improvement on short horizons (1 to 2 hours) in average 30% over a persistence reference model, and in average 15% on horizons up to 42 hours. Applying the NWPs an improvement around 39 % is achieved in average for short horizons and around 32% in average for longer horizons. The method can furthermore be applied to monitor and check the performance of solar thermal collectors.

References

- P. Bacher, H. Madsen, and H. A. Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83(10):1772–1783, 2009. ISSN 0038092x.
- DMI. Danish Meteorological Institute, DMI-HIRLAM-S05, 2011. URL http://www.dmi.dk/eng/index/research_and_development/dmi-hirlam-2009.htm.

- H. Madsen. *Time Series Analysis*. Chapman & Hall, 2007.
- H. Madsen, P. Pinson, G. Kariniotakis, H. A. Nielsen, and T. S. Nielsen. Standardizing the performance evaluation of shortterm wind power prediction models. *Wind Engineering*, 29(6):475, 2005. ISSN 0309524x.
- B. Perers. An improved dynamic solar collector test method for determination of non-linear optical and thermal characteristics with multiple regression. *Solar Energy*, 59(4-6):163–178, 1997. ISSN 0038092x.
- B. Perers, S. Furbo, J. Fan, E. Andersen, and Z. Chen. Solar combisystems with forecast control to increase the solar fraction and lower the auxiliary energy cost. In *ISES Solar World Congress 2011 Proceedings*, page , 2011. ISBN 978-3-9814659-0-7. Presented at: ISES Solar World Congress, SWC ; 30 : Kassel, Germany, 2011.

P A P E R E

Models of the heat dynamics of solar collectors for performance testing

Authors:

Peder Bacher¹, Henrik Madsen¹, and Bengt Perers²

In proceedings:

ISES Solar World Congress 2011.

¹DTU Informatics, Richard Pedersens Plads, Building 321, DK-2800 Lyngby, Denmark

²DTU Civil Engineering, Brovej, Building 118, DK-2800 Lyngby, Denmark

Abstract

The need for fast and accurate performance testing of solar collectors is increasing. This paper describes a new technique for performance testing which is based on non-linear continuous time models of the heat dynamics of the collector. It is shown that all important performance parameters can be accurately estimated with measurements from a single day. The estimated parameters are compared with results from standardized test methods (Fischer et al., 2004).

Modelling the dynamics of the collector is carried out using stochastic differential equations, which is a well proven efficient method to obtain accurate estimates of parameters in physical models. The applied method is described by Kristensen et al. (2004) and implemented in the software CTSM¹. Examples of successful applications of the method includes modelling the of the heat dynamics of integrated photo-voltaic modules (Friling et al., 2009) and modelling of the heat dynamics of buildings (Madsen and Holst, 1995).

Measurements obtained at a test site in Denmark during the spring 2010 are used for the modelling. The tested collector is a single glazed large area flat plate collector with selective absorber and Teflon anti convection layer. The test rig is described in Fan et al. (2009).

The modelling technique provides uncertainty estimates such as confidence intervals for the parameters, and furthermore enables statistical validation of the results. Such tests can also facilitate procedures for selecting the best model to use, which is a very non-trivial task.

E.1 Introduction

This paper presents a new statistical approach for modelling the heat dynamics of a solar thermal collector. The applied modelling technique facilitates application of detailed models on data sampled with a relatively high sample rate. It is demonstrated that this allows for parameter estimation with high accuracy to be achieved with measurements from a single day. In the present study 2 seconds values averaged to 30 seconds values are used. Conventional non-dynamical models - by some called pseudo-dynamical models - of solar collectors cannot use such a high sample rate due to auto-correlation of the errors caused by non-modelled dynamical effects. The applied models are based on stochastic differential equations (SDEs), which gives the possibility to combine physical

¹www2.imm.dtu.dk/~ctsm/

and data-driven statistical modelling. Such models are called grey-box models. A very strong feature of grey-box models is that they provide the possibility to estimate hidden state variables, i.e. variables in the model which are not measured. This allows using the same data for fitting models, with which the system is lumped differently, i.e. the physical model of the system can either be a single-state or a multi-state lumped model, which can be required for different types of collectors. Furthermore the modelling technique facilitates application of statistical tests to determine which model is most suitable for the given data. This is important for model identification and the approach is demonstrated in the paper. The modelling is carried out based on measurements from a period of 9 consecutive days in the beginning of May 2010. None of the days could have been used for stationary testing that is still the most common test method for solar collectors. Stationary testing requires perfect stable clear weather around noon. The measurements were performed on a single glazed large area flat plate collector with selective absorber and Teflon anti convection layer. The collector was not brand new, but has been in operation for 15 years, which affects the parameter values compared to today's products of the similar design. The results from the grey-box models are compared with results from the standardized EN 12975 Quasi Dynamic Test Method (CEN, European committee for standardization, 2006), which is based on multiple linear regression (MLR) modelling, to see if the estimation results matches current test standards. Finally, a thorough discussion and perspectives of the technique are given.

The paper is arranged as follows. The next section is a presentation of the theory of the applied grey-box models, with a simple example. This is followed by a section with a description of the MLR models used and thereafter a section with a description of all the considered grey-box models. Then the results are presented, and finally a discussion and perspective is given before the paper ends with a conclusion.

Nomenclature

The same notation as in Perers (1997) are used as widely as possible.

Collector model parameters:

$F'(\tau\alpha)_{\text{en}}$	Zero loss efficiency for direct radiation at normal incidence
$K_{\tau\alpha b}(\theta)$	Incidence angle modifier for direct radiation
$K_{\tau\alpha d}$	Incidence angle modifier for diffuse radiation
$F'U_0$	Heat loss coefficient at $(T_a - T_f) = 0$, $[\text{W}/(\text{m}^2\text{K})]$.
$F'U_1$	Temperature dependence of the heat loss coefficient, $[\text{W}/(\text{m}^2\text{K}^2)]$.
$F'U_w$	Wind dependence of the heat loss coefficient, $[\text{Ws}/(\text{m}^3\text{K})]$.
$(mC)_e$	Effective thermal capacitance including piping for the collector, $[\text{J}/(\text{m}^2\text{K})]$.
C_f	Fluid thermal capacitance, $[\text{J}/(\text{m}^2\text{K})]$.
C_m	Collector thermal capacitance, $[\text{J}/(\text{m}^2\text{K})]$.
U_{fa}	Heat transmission coefficient from fluid to ambient, $[\text{J}/(\text{Km}^2)]$.
U_{fm}	heat transmission coefficient from fluid to module, $[\text{J}/(\text{Km}^2)]$.
U_{ma}	heat transmission coefficient from module to ambient, $[\text{J}/(\text{Km}^2)]$.
n_c	Number of compartments

Measured variables:

G_d	Diffuse radiation onto the collector plane, $[\text{W}/\text{m}^2]$.
G_b	Direct radiation onto the collector plane, $[\text{W}/\text{m}^2]$.
T_a	Ambient air temperature near the collector, $[\text{°C}]$.
T_o	Outlet temperature of the collector, $[\text{°C}]$.
T_i	Temperature of the inlet to the collector, $[\text{°C}]$.
Q_f	Flow of the fluid per square meter of collector, $[\text{l}/(\text{sm}^2)]$.
θ	incidence angle for the direct solar radiation onto the collector plane, [radians].
w	Wind speed, $[\text{m}/\text{s}]$.

Derived variables etc.:

T_f	Average temperature of the collector fluid, $[\text{°C}]$.
T_m	Average temperature of the collector, $[\text{°C}]$.
q_u	Collector power output, $[\text{W}/\text{m}^2]$.
c_f	Specific heat capacity of the fluid, $[\text{J}/(\text{IK})]$.

E.2 Grey-box models of a dynamic system

A grey-box model is established using a combination of prior physical knowledge and statistics, i.e. information embedded in data Kristensen et al. (2004). The prior physical knowledge is formulated by a set of non-linear stochastic differential equations (SDEs), also called a stochastic non-linear state-space model in continuous time. The equations describe a lumped model of the heat dynamics of the system.

The output of the solar collector is calculated by

$$q_u = c_f Q_f (T_o - T_i) \quad (\text{E.1})$$

where T_o is the outlet temperature and T_i is the inlet temperature of the fluid. The output q_u is power output per square meter of collector aperture area and Q_f is flow per the same area. From Perers (1997) it is known that the output of a standard flat plate collector in first order accuracy level can be described by the heat balance

$$\begin{aligned} c_f Q_f (T_o - T_i) = & F'(\tau\alpha)_{\text{en}} K_{\tau\alpha b}(\theta) G_b + F'(\tau\alpha)_{\text{en}} K_{\tau\alpha d} G_d \\ & - F' U_0 (T_f - T_a) - (mC)_e \frac{dT_f}{dt} \end{aligned} \quad (\text{E.2})$$

For explanation of the symbols, see page 128. A linear temperature profile through the collector is applied by modelling the fluid temperature as a simple average

$$T_f = \frac{T_o + T_i}{2} \quad (\text{E.3})$$

The differential of the fluid temperature can then be written as

$$\frac{dT_f}{dt} = \frac{1}{2} \frac{dT_o}{dt} + \frac{1}{2} \frac{dT_i}{dt} \quad (\text{E.4})$$

which for a constant inlet temperature is

$$\frac{dT_f}{dt} = \frac{1}{2} \frac{dT_o}{dt} \quad (\text{E.5})$$

This substitution, together with the addition of a noise term, is used to form the SDE

$$\begin{aligned} dT_o = & \left(F' U_0 (T_a - T_f) + c_f Q_f (T_i - T_o) \right. \\ & \left. + F'(\tau\alpha)_{\text{en}} K_{\tau\alpha b}(\theta) G_b + F'(\tau\alpha)_{\text{en}} K_{\tau\alpha d} G_d \right) \frac{2}{(mC)_e} dt + \sigma d\omega \end{aligned} \quad (\text{E.6})$$

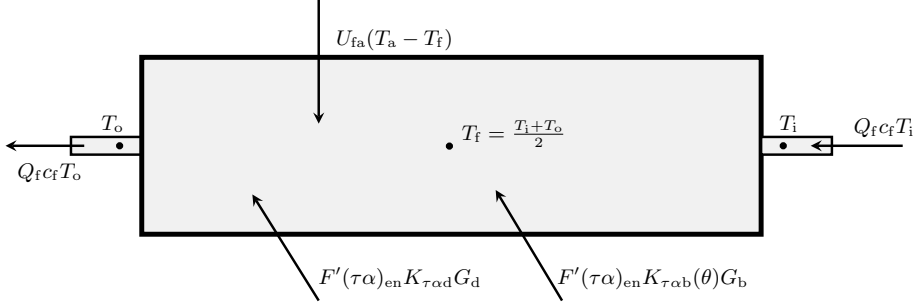


Figure E.1: Diagram of *ToComp1* illustrating all the energy flows included in the model.

which describes the heat dynamics for the collector in the simplest grey-box model considered in the paper. It is denoted as *ToComp1*. In grey-box terminology this is called the system equation of the state-space model. The noise term $\sigma d\omega$ is called the system noise and consist of increments of $\{\omega\}$, which is a standard Wiener process, and σ^2 , which is the incremental variance of the Wiener process. In this model the collector is lumped into one single part and the state variable is the outlet temperature T_o . An illustration of the model is found in Figure E.1.

The physical model part is coupled with the data-driven model part with which the information embedded in observed data is used for parameter estimation. The data-driven part in the considered example is represented by the discrete time measurement equation

$$Y_k = T_{ok} + e_k \quad (\text{E.7})$$

where k is the point in time t_k of a measurement, Y_k is the measured outlet temperature, and e_k is the measurement error, which is assumed to be a Gaussian white noise process with variance σ^2 . This assumption - plus the assumption that W is a Wiener process - enables evaluation and tests of the performance of the model, since such tests can show if the physical model is consistent with the observed heat dynamics of the collector.

E.2.1 Maximum likelihood estimation of parameters

Given a grey-box model, as described above, maximum likelihood estimates of the parameters can be obtained. Let the N observations be represented by

$$\mathcal{Y}_N = [Y_N, Y_{N-1}, \dots, Y_1, Y_0] \quad (\text{E.8})$$

then the likelihood function is the joint probability density

$$L(\theta; \mathcal{Y}_N) = \left(\prod_{k=1}^N p(Y_k | \mathcal{Y}_{k-1}, \theta) \right) p(Y_0 | \theta) \quad (\text{E.9})$$

where $p(Y_k | \mathcal{Y}_{k-1}, \theta)$ is a conditional density denoting the probability of observing Y_k given the previous observations and the parameters θ , and where $p(Y_0 | \theta)$ is a parameterization of the starting conditions Kristensen et al. (2004). The maximum likelihood estimates of the parameters are then found by

$$\hat{\theta} = \arg \max_{\theta} \{L(\theta; \mathcal{Y}_N)\} \quad (\text{E.10})$$

Due to the previously mentioned assumptions about the system and measurement noise, it follows that the conditional densities in Equation (E.10) can be well approximated by Gaussian densities. Hence an extended Kalman filter can be used to calculate the likelihood function, and an optimization algorithm can be applied to maximize it, thereby calculating the maximum likelihood estimates, see Kristensen et al. (2004) for a detailed discussion. This is implemented in the computer software CTSM, which has been used for carrying out the parameter estimation. See more about the methods and software at ² and in Kristensen and Madsen (2003).

E.3 Experimental setup and data

The experiments are described by Fan et al. (2009) and were carried out in the spring of 2010. The measurements were obtained with a 2 seconds sample interval. For the present study models are identified for both 30 seconds and 10 minutes average values. The data resampled to 10 minute average values is plotted in Figure E.2. Only time points where the angle of incidence is lower than 84 degrees are used. For the parameter estimation it is important to acquire a period, for which the input signals are as uncorrelated as possible and cover the typical range of operation. Periods with full cloud cover are not feasible, since there is not enough variation in the direct radiation and in periods with no cloud cover the radiation and the module temperature is highly correlated. Hence days with varying cloud cover are most appropriate and these days are the most common in most locations where people traditionally live.

²www.imm.dtu.dk/~ctsm

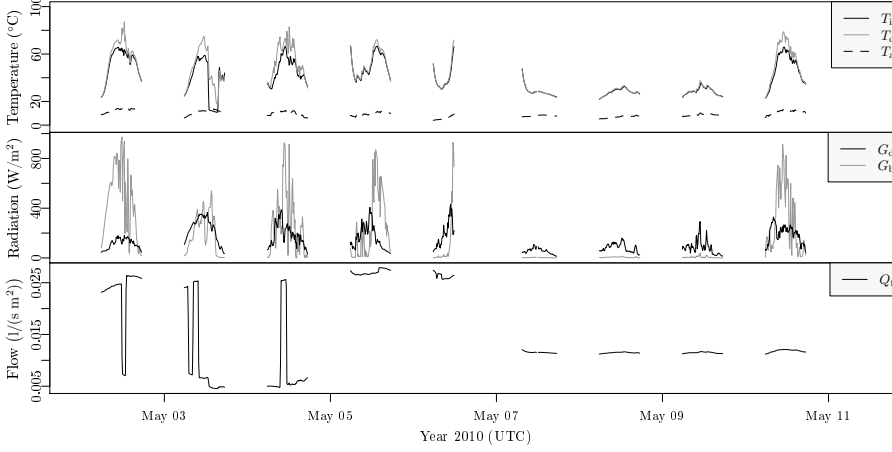


Figure E.2: The data as 10 minutes averaged values. The upper plot is the measured temperatures, the middle plot is the diffuse and direct (beam) radiation, and the lowest plot is the fluid flow.

E.4 Multiple linear regression models

The EN 12975 Quasi Dynamic Test Method (CEN, European committee for standardization, 2006) is applied to have a reference for the results from the new proposed method. The method is based on multiple linear regression (MLR) modelling, where down to 5 minutes average values are recommended. The data was resampled to 10 minutes averages, which for all 9 days gives 593 time points. MLR modelling with 5 minutes averages was tried and the results were only marginally different. The following model structure is applied

$$q_t^u = F'(\tau\alpha)_{\text{en}} K_{\tau\alpha b}(\theta) G_t^b + F'(\tau\alpha)_{\text{en}} K_{\tau\alpha d} G_t^d + F'U_0 \Delta T_t + F'U_1 \Delta T_t^2 - F'U_w \Delta T_t w_t - (mC)_e \frac{dT_t^f}{dt} + e_t \quad (\text{E.11})$$

where $\Delta T_t = T_t^a - T_t^f$. Three models are fitted: *MLR1* without $F'U_1$ and $F'U_w$, *MLR2* without $F'U_1$, and *MLR3* with all inputs.

E.5 Applied grey-box models

This section gives an overview of the applied grey-box models and the parts of the data on which the parameter estimation was carried out. First the single

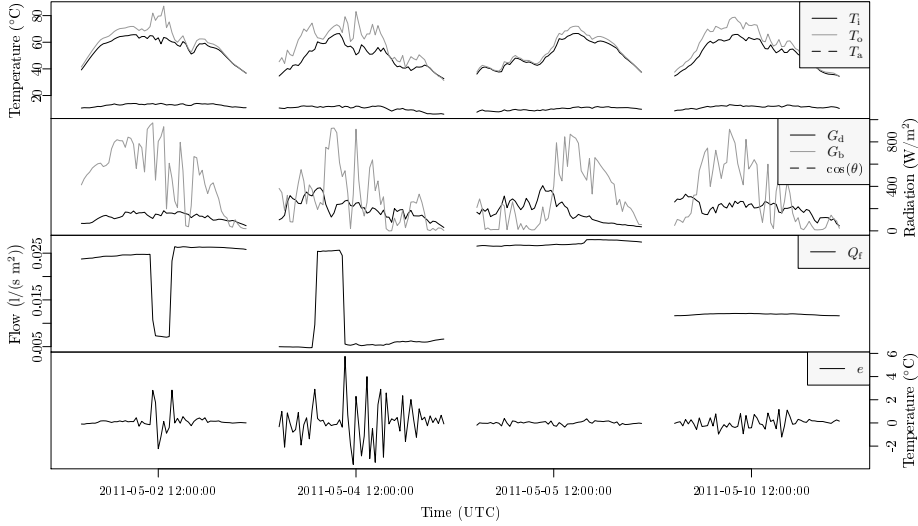


Figure E.3: The 10 minutes averaged values from days where the model *ToComp1* is fitted. The upper plot is of the inlet-, outlet-, and the ambient temperature. Below this is shown a plot of the direct- and diffuse solar radiation, followed by a plot of the fluid flow. The lowest plot is the residuals from the fit from each day, this is referred to in a later in the paper.

state grey-box model *ToComp1*, described in Section E.2 was fitted to 10 minutes average values on the days with varying cloud cover. This data is plotted in Figure E.3. The model was fitted to data from each day separately and finally to all the data from four days pooled together. In addition to the *ToComp1* model four other grey-box models have been fitted to the data from the 10th of May resampled to 30 second average values. This gives $N = 1413$ data points, which are plotted in Figure E.4. The additional four models are expanded as more detailed versions of *ToComp1*. There are two ways to expand the model: either more inputs (explanatory variables) can be used, or - since the models are lumped models - a better representation can maybe be achieved by lumping the system into more parts (also called compartments, states, zones, or nodes). The latter approach is considered in the following. The first two expanded models are made more detailed by lumping the collector into more than one compartment in the flow direction of the collector fluid, such that the temperature of the collector is represented by two or more temperature state variables. This allows for a better representation of the temperature differences between the inlet - the cold side when solar radiation level is high - and the outlet of the collector. Furthermore this also gives a better description of the delay introduced since it takes time for the fluid to flow through the collector. For the current setup

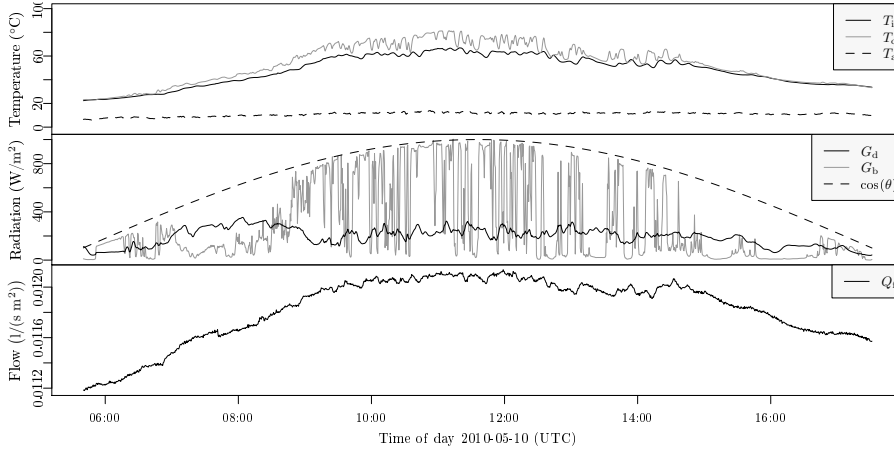


Figure E.4: The 30 seconds averaged data for which the modelling is applied. The upper plot is of the inlet-, outlet-, and the ambient temperature. Below this is a plot of the direct- and diffuse solar radiation, followed by the plot of the fluid flow.

and the flow of the 10'th of May, this is around 1 minute. These two models are denoted by *ToComp2* and *ToComp3*. The third expanded model is denoted by *ToTmComp1* and in this model the collector is lumped in two parts: one representing the fluid and one representing the solid part of the collector. This is a better description of the system, in which the solar radiation first heats up the collector which then heats up the fluid. Finally, the fourth expanded model *TmToComp2* is a combination of the two approaches, where the collector is first divided in two parts - one for the fluid and one for the collector - which then each are divided into two compartments in the flow direction of the fluid.

E.5.1 Models with multiple compartments in the flow direction

The *ToComp1* model can be expanded to a n_c compartment model with the system equations

$$\begin{aligned}
 dT_{o1} &= \left(F'U_0(T_a - T_{f1}) + n_c c_f Q_f (T_i - T_{o1}) + F'(\tau\alpha)_{en} K_{\tau\alpha b}(\theta) G_b \right. \\
 &\quad \left. + F'(\tau\alpha)_{en} K_{\tau\alpha d} G_d \right) \frac{2}{(mC)_e} dt + \sigma_1 d\omega_1 \\
 dT_{o2} &= \left(F'U_0(T_a - T_{f2}) + n_c c_f Q_f (T_{o1} - T_{o2}) + F'(\tau\alpha)_{en} K_{\tau\alpha b}(\theta) G_b \right. \\
 &\quad \left. + F'(\tau\alpha)_{en} K_{\tau\alpha d} G_d \right) \frac{2}{(mC)_e} dt + \sigma_2 d\omega_2 \\
 &\vdots \\
 dT_{on_c} &= \left(F'U_0(T_a - T_{fn_c}) + n_c c_f Q_f (T_{o(n_c-1)} - T_{on_c}) + F'(\tau\alpha)_{en} K_{\tau\alpha b}(\theta) G_b \right. \\
 &\quad \left. + F'(\tau\alpha)_{en} K_{\tau\alpha d} G_d \right) \frac{2}{(mC)_e} dt + \sigma_{n_c} d\omega_{n_c}
 \end{aligned} \tag{E.12}$$

where n_c is the number of compartments. The accompanying measurement equation is

$$Y_k = T_{on_c k} + e_k \tag{E.13}$$

Two models of this type are fitted to the data: *ToComp2* with two compartments, and *ToComp3* with three compartments. A diagram illustrating *ToComp2* is shown in Figure E.5

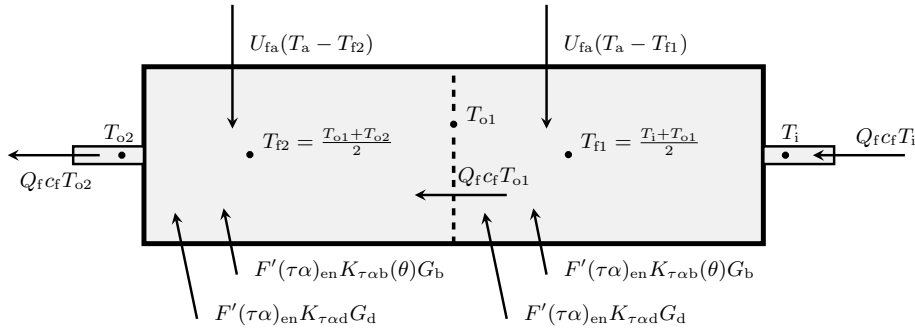


Figure E.5: The *ToComp2* model with two compartments in the flow direction of the fluid.

E.5.2 Models divided into a collector and a fluid part

The *TmToComp1* model illustrated in Figure E.6, where the panel is divided into two parts, has the system equation

$$\begin{aligned} dT_m &= \left(F'(\tau\alpha)_{\text{en}} K_{\tau\alpha b}(\theta) G_b + F'(\tau\alpha)_{\text{en}} K_{\tau\alpha d} G_d + U_{\text{fm}}(T_f - T_m) \right. \\ &\quad \left. + U_{\text{ma}}(T_a - T_m) \right) \frac{1}{(mC)_e} dt + \sigma_m d\omega_m \\ dT_o &= \left(U_{\text{fm}}(T_m - T_f) + c_f Q_f(T_i - T_o) \right) \frac{2}{(mC)_e} dt + \sigma_o d\omega_o \end{aligned} \quad (\text{E.14})$$

It is seen that the solar radiation enters the collector part, which then heats up the fluid by conduction. Of the considered models the most detailed model is *TmToComp2*, in which the collector is both divided into two parts and 2 compartments in the fluid flow direction for each part. The following system equations is formulated for a model with two parts having each n_c compartments

$$\begin{aligned} dT_{m1} &= \left(F'(\tau\alpha)_{\text{en}} K_{\tau\alpha b}(\theta) G_b + F'(\tau\alpha)_{\text{en}} K_{\tau\alpha d} G_d + U_{\text{fm}}(T_{f1} - T_{m1}) \right. \\ &\quad \left. + U_{\text{ma}}(T_a - T_{m1}) \right) \frac{1}{(mC)_e} dt + \sigma_{m1} d\omega_{m1} \\ dT_{o1} &= \left(U_{\text{fm}}(T_{m1} - T_{f1}) + n_c c_f Q_f(T_i - T_{o1}) \right) \frac{2}{(mC)_e} dt + \sigma_{o1} d\omega_{o1} \\ dT_{m2} &= \left(F'(\tau\alpha)_{\text{en}} K_{\tau\alpha b}(\theta) G_b + F'(\tau\alpha)_{\text{en}} K_{\tau\alpha d} G_d + U_{\text{fm}}(T_{f2} - T_{m2}) \right. \\ &\quad \left. + U_{\text{ma}}(T_a - T_{m2}) \right) \frac{1}{(mC)_e} dt + \sigma_{m2} d\omega_{m2} \\ dT_{o2} &= \left(U_{\text{fm}}(T_{m2} - T_{f2}) + n_c c_f Q_f(T_{o1} - T_{o2}) \right) \frac{2}{(mC)_e} dt + \sigma_{o2} d\omega_{o2} \\ &\vdots \\ dT_{mn_c} &= \left(F'(\tau\alpha)_{\text{en}} K_{\tau\alpha b}(\theta) G_b + F'(\tau\alpha)_{\text{en}} K_{\tau\alpha d} G_d + U_{\text{fm}}(T_{fn_c} - T_{mn_c}) \right. \\ &\quad \left. + U_{\text{ma}}(T_a - T_{mn_c}) \right) \frac{1}{(mC)_e} dt + \sigma_{mn_c} d\omega_{mn_c} \\ dT_{on_c} &= \left(U_{\text{fm}}(T_{mn_c} - T_{fn_c}) + n_c c_f Q_f(T_{o1} - T_{on_c}) \right) \frac{2}{(mC)_e} dt + \sigma_{on_c} d\omega_{on_c} \end{aligned} \quad (\text{E.15})$$

i.e. the *TmToComp2* model has $n_c = 2$.

E.6 Results

In this section the results of the parameter estimation with the described models are presented. Firstly, the results from the traditional MLR models fitted on the entire data set of 10 minutes values is presented, secondly from grey-box model *ToComp1* fitted on individual days of 10 minutes values, and finally all grey-box models fitted on 30 seconds values from the 10'th of May. The parameter estimates together with the their standard deviation are presented in tables, and time series of the residuals together with other relevant error measures are plotted. A short outline of the model identification carried out is also provided.

E.6.1 MLR models

The parameter estimates are listed in Table E.1. The estimates are clearly within the typical range for this type of collector, see Perers (1993) and Solar Keymark homepage (Solar Keymark, 2011). The collector under test has been in operation for 15 years, this affects the parameter values compared to todays products. The standard deviations show that the parameters are very accurately determined. The only non-significant term are $F'U_1$ in *MLR3*, which leads the conclusion that *MLR2* is the most appropriate model of the three. For evaluation of the model fit the measured collector output versus the predicted is plotted in Figure E.7. It is seen that the measured output is predicted very well, although it does seems like the variance increase slightly with the output. This is most likely due to the periods with low flow rate for some of the days. Furthermore the inlet temperature variation range is not fully as high as specified in the standard for the selected days.

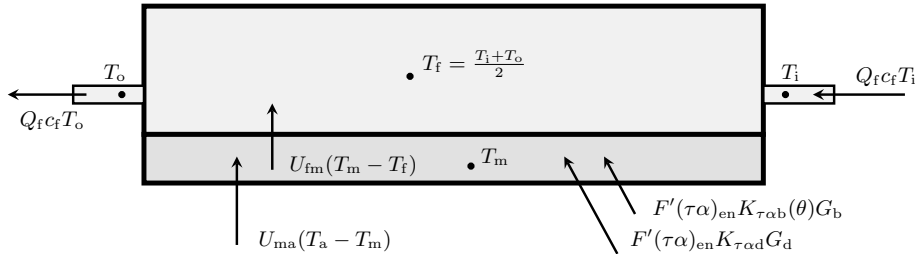


Figure E.6: Diagram illustrating the *TmToComp1* model. The collector is divided into a part representing the fluid and another part representing the collector.

Parameter	<i>MLR1</i>		<i>MLR2</i>		<i>MLR3</i>		Units
$F'(\tau\alpha)_{\text{en}}$	0.737	(0.0031)	0.741	(0.0030)	0.746	(0.0043)	
b_0	0.166		0.172		0.175		
$K_{\tau\alpha d}$	0.891		0.904		0.895		
$F'U_0$	2.18	(0.45)	2.13	(0.045)	2.02	(0.082)	$[\text{W}/(\text{m}^2\text{K})]$
$F'U_1$					0.0031*	(0.0020)	$[\text{W}/(\text{m}^2\text{K})]$
$F'U_w$			0.192	(0.034)	0.179	(0.035)	$[\text{W}/(\text{m}^2\text{K})]$
$(mC)_e$	4699	(130)	4751	(127)	4788	(129)	$[\text{J}/(\text{m}^2\text{K})]$

Table E.1: Parameter estimates with MLR models. The standard deviation of the estimate is in parenthesis to the right of the estimated value. Insignificant terms are marked with *.

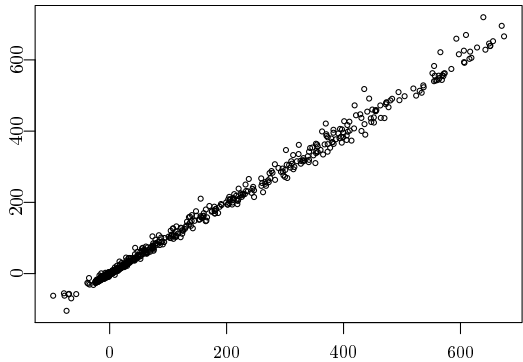


Figure E.7: Measured versus the predicted collector output from *MLR2*.

E.6.2 ToComp1 fitted to 10 minutes values

The single state grey-box model defined in Equation (E.6) is fitted to both 10 minutes values from four separate days and all four pooled together. The estimated parameter are listed in Table E.2. Clearly the parameter estimates matches the estimates from the MLR models quite well considering the standard deviations, especially the parameters $F'(\tau\alpha)_{\text{en}}$ and $F'U_0$, which are the most important parameters for evaluation of the collector performance. A very apparent deviation of the results between the days is that the lowest uncertainty is found on the 5'th of May. This is not a surprise considering a plot of the residuals, which is shown in the lowest plot of Figure E.3. The level of the residuals from the fit to this day are smaller than for the other days, and this is apparently due to the level of the fluid flow, which is plotted above the residuals in the figure. The parameter estimates based on the four days pooled together seems like a compromise between the estimates from the single days.

	2011-05-02		2011-05-04		2011-05-05	
$F'(\tau\alpha)_{\text{en}}$	0.755	(0.032)	0.785	(0.032)	0.746	(0.0086)
b_0	0.204	(0.039)	0.201	(0.046)	0.18	(0.017)
$K_{\tau\alpha\text{d}}$	0.903	(0.42)	0.857	(0.11)	0.819	(0.027)
$F'U_0$	2.07	(1.1)	2.4	(0.35)	1.73	(0.13)
$(mC)_e$	6050	(1060)	6200	(1130)	5040	(279)
	2011-05-10		Pooled			
$F'(\tau\alpha)_{\text{en}}$	0.758	(0.014)	0.763	(0.011)		
b_0	0.182	(0.023)	0.195	(0.020)		
$K_{\tau\alpha\text{d}}$	0.867	(0.049)	0.839	(0.034)		
$F'U_0$	2.16	(0.26)	2.05	(0.18)		
$(mC)_e$	5020	(92)	5666	(638)		

Table E.2: Parameter estimates from *ToComp1* fitted to 10 minutes values from single days and all four days pooled. The standard deviation is given in parenthesis to the right of the estimate.

E.6.3 Grey-box models fitted to 30 seconds values

The five grey-box models described are fitted to the data from the 10'th of May resampled to 30 seconds averages. The parameter estimates are listed in Table E.3. First, it is noticed that the parameters of the three models prefixed with *To* are not representing the same physical entities as they do in the models prefixed with *ToTm*, since the collector is lumped differently in the models. The increase of the value of $F'(\tau\alpha)_{\text{en}}$ from *To* to *ToTm* models is found to be consistent with the physical representation, since the reference temperature is closer to the absorber surface. This means that the estimated optical parameter for the *ToTm* models is rather $\tau\alpha$. The value F' is in the range of 0.95 for this collector design, which leads to an estimate of $F'(\tau\alpha)_{\text{en}}$ to 0.752 for *ToTmComp2*.

Plots of the residual series from each model are shown in Figure E.8. Clearly the level of the residuals decrease from the upper to the lower plot and the highest errors occur when a cloud passes by and the level of direct solar radiation shifts rapidly. The decreased level of the variation of the residuals indicates that the more detailed models are better. To verify this, statistical likelihood-ratio tests is applied as described by Bacher and Madsen (2011). The log-likelihood of the fit for each model is listed in Table E.4, together with the number of parameters, and the p -value of tests for model expansion. The tests for expansion is a model versus the expanded model a single step to the right in the table, except for *nl2TmToComp1*, from which the expansion is from *nl2ToComp2*. The results of the tests are very clear, all expansions are significant. Hence it is concluded

140 Models of the heat dynamics of solar collectors for performance testing

Prm.	ToComp1		ToComp2		ToComp3	
$F'(\tau\alpha)_{\text{en}}$	0.767	(0.0036)	0.751	(0.0027)	0.743	(0.0015)
b_0	0.172	(0.0063)	0.177	(0.0017)	0.18	(0.00044)
$K_{\tau\alpha d}$	0.942	(0.015)	0.933	(0.0042)	0.931	(0.002)
U_{fa}	2.55	(0.076)	2.31	(0.049)	2.2	(0.023)
U_{fm}						
U_{ma}						
C_f	6960	(80)	8020	(17)	8580	(36)
C_m						
Prm.	TmToComp1		TmToComp2			
$F'(\tau\alpha)_{\text{en}}$	0.816	(0.0025)	0.792	(0.00096)		
b_0	0.188	(0.0038)	0.189	(0.00067)		
$K_{\tau\alpha d}$	0.929	(0.008)	0.927	(0.0021)		
U_{fa}						
U_{fm}	49.8	(2.5)	83.7	(0.83)		
U_{ma}	2.37	(0.042)	2.22	(0.016)		
C_f	3750	(114)	3390	(54)		
C_m	962	(64)	1690	(22)		

Table E.3: The parameter estimates from the grey-box models fitted to 30 seconds values from the 10th of May. Note that the parameters represent different physical entities from the three first model (prefixed with *To*) to the last two models (prefixed with *TmTo*) and therefore cannot be directly compared. For each estimate the standard deviation is given in parenthesis to the right of the estimate.

that *nl2TmToComp2* is the most suitable model of these five models and that it might very well be, that the model could be further expanded. Finally, the auto-correlation function (ACF) and the cumulated periodogram (CP) Madsen (2007) of the residuals are shown in Figure E.9. The dashed blue lines indicate 95% confidence intervals for a white noise. According to theory, presented in Section E.2, then if the residual series are white noise this indicates that the grey-box model is consistent with the observed heat dynamics of the collector. From the ACF and CP it is seen that the residuals are close to having white noise properties. Interestingly it is seen that residuals from *ToComp1* are more white noise like than the residuals from *TmToComp1*. It is found that this is caused by a low signal to noise ratio in the residuals for *ToTmComp1*, i.e. the dominating errors are on a high frequency which have characteristics like white noise. As the detailed models includes the faster dynamics, the high frequency errors are decreased, and they do not “hide” the remaining signal components in the residuals. However for the most detailed model almost all the systematic variation in the data is described.

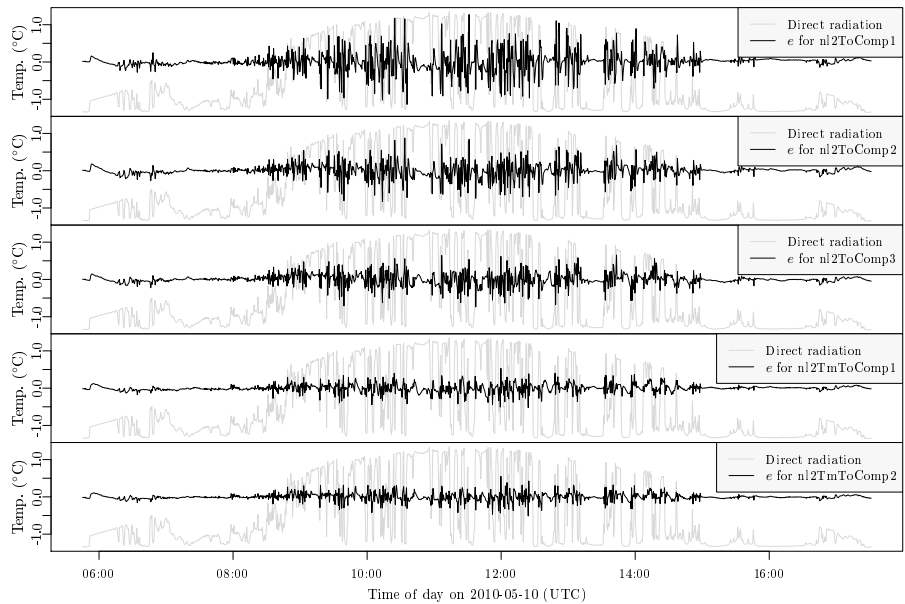


Figure E.8: The residual series from the grey-box models fitted to 30 seconds values from the 10'th of May. The greyed series are the direct solar radiation.

	nl2ToComp1	nl2ToComp2	nl2ToComp3
Log-likelihood	-35.51	454.8	661
Number of prm.	9	12	14
p -value		≈ 0	≈ 0
	nl2TmToComp1	nl2TmToComp2	
Log-likelihood	1185	1307	
Number of prm.	13	18	
p -value	≈ 0	≈ 0	

Table E.4: Log-likelihood, number of parameters, and p -value of likelihood-ratio tests for model expansion for each of the grey-box models.

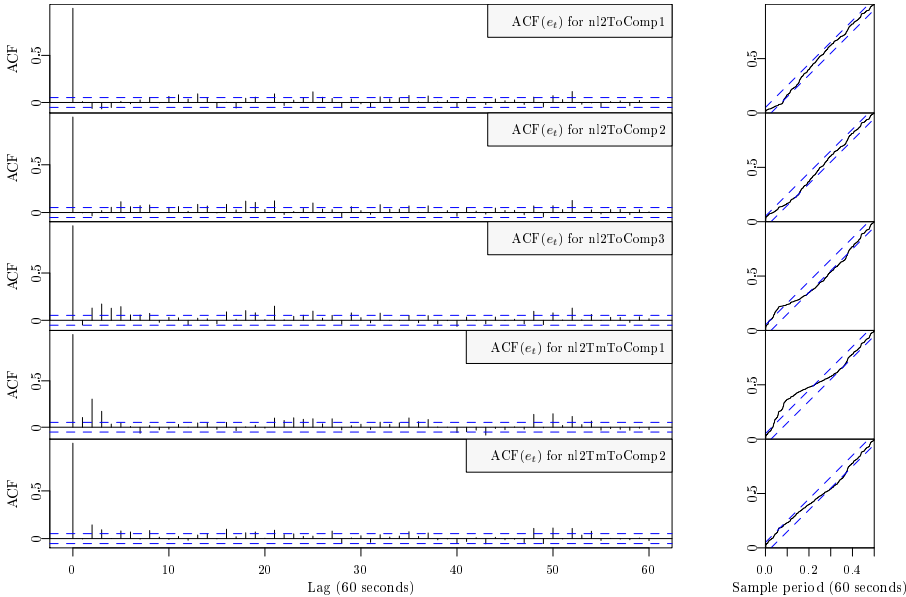


Figure E.9: Plots of the auto-correlation function (ACF) and the cumulated periodogram of the residuals from each of the grey-box models fitted on 30 seconds values.

E.7 Discussion and applications

In general the results from the MLR models and the grey-box are found to match well, but it is noted that the result the from grey-box model *ToComp1* fitted on separate days - which have different conditions, especially in the fluid flow level - gives some variation to the estimates. Therefore it is concluded that attention has to be put on the experimental design in order to ensure stable and accurate parameter estimation for collector testing with grey-box models. Regarding the more detailed grey-box models fitted to 30 seconds values, it is found that since the likelihood is not saturated, i.e. the likelihood-ratio tests are very significant, further expansion of the *TmToComp2* is still possible. From the plots of the residuals in Figure E.8, it is seen that the error level certainly is highest just after the direct radiation shifts its level very rapidly, and it is this effect that seems to be improved as the more detailed models are used. Hence the main improvement from the one-state model *ToComp1* to the multi-state models are in the description of the fast dynamics, which includes the delay caused by movement of the fluid through the collector, e.g. when the direct radiation shifts from a high to a low level, the fluid passing out of the collector are still hot for some time. This also indicates the importance of the experiment design, since for dynamic condition the frequency, with which the system is excited, affect which grey-box model is optimal. For example if the direct radiation varies with a lower frequency, a simpler model might be in favour over more complex models, whereas for variation with a higher frequency the inclusion of the fast dynamics are more important. Therefore if the main excitation of the system, i.e. the direct radiation, can be controlled, it will be possible to achieve fast and accurate parameter estimation. This could be carried out with a simple shadowing device, which should be controlled with PRBS signal to gain maximum information of the heat dynamics of the system (Madsen and Holst, 1995). Higher accuracy can also be achieved with more systematic variation of the inlet temperature, this also applies for the MLR modelling. The right experiment design will allow inclusion of night measurements - which will improve the separation of heat loss and radiation effects and thereby more accurate estimation - and furthermore allow for inclusion of more effects, such as wind and non-linear radiation effects between the collector and the surroundings. Finally, dividing the collector into more parts, e.g. one representing the fluid, one representing the metal, and one representing the surrounding collector body could be tried.

E.7.1 Applications

The most apparent application of grey-box modelling of the heat dynamics of solar collectors are for the development of fast and accurate performance testing,

especially for some types of collectors multi-state models are needed to obtain a required level of accuracy. Especially vacuum tube collectors of dewar type can have an extra time delay due to the high thermal resistance between the heat transfer fluid and absorber surface that is not fully taken up by the present collector model used for performance testing. The new approach described here, particularly with the *TmTo* models, has the potential to deal with this in an accurate way. Additional applications include optimization of operation with model predictive control, which the grey-box models are perfectly suited for. Especially larger solar thermal plants might be able to gain much in performance by applying grey-box modelling and model predictive control.

E.8 Conclusion

Successful modelling of the heat dynamics of a solar collector with grey-box models has been carried out. The results have been compared to the EN-standard MLR modelling and they are in agreement. It is shown that high accuracy parameter estimates were obtained with measurements from a single day resampled to 30 seconds average values. This will enable lowering of testing time significantly compared to current test methods. Highly detailed models of the heat dynamics of the solar collector can be applied, which can be useful for many types of collectors. It is found that the conditions under which the experiment was carried out influence the parameter estimates. Therefore it is concluded that experiment design is the key to achievement of fast, reliable and high accuracy collector testing methods with grey-box models. Experiments with PRBS variation of direct radiation with shadowing device should be carried out to obtain higher accuracy and reproducibility of the results, and finally models with more explanatory variables, such as wind and long-wave radiation should be further elaborated.

References

- P. Bacher and H. Madsen. Identifying suitable models for the heat dynamics of buildings. *Energy & Buildings*, 43(7):1511–1522, 2011. ISSN 03787788. doi: 10.1016/j.enbuild.2011.02.005.
- CEN, European committee for standardization. En 12975-2:2006, thermal solar systems and components - collectors - part 2: Test methods, 2006.
- J. Fan, Z. Chen, S. Furbo, B. Perers, and B. Karlsson. Efficiency and lifetime of

- solar collectors for solar heating plants. Proceedings of the ISES Solar World Congress 2009: Renewable Energy Shaping Our Future, 2009.
- S. Fischer, W. Heidemann, H. Müller-Steinhagen, B. Perers, P. Bergquist, and B. Hellström. Collector test method under quasi-dynamic conditions according to the european standard en 12975-2. *Solar Energy*, 76(1-3):117–123, 2004. ISSN 0038092x. doi: 10.1016/j.solener.2003.07.021.
- N. Friling, M. J. Jiménez, H. Bloem, and H. Madsen. Modelling the heat dynamics of building integrated and ventilated photovoltaic modules. *Energy and Buildings*, 41(10):1051–1057, 2009. ISSN 03787788.
- N. R. Kristensen and H. Madsen. Continuous time stochastic modelling, CTSM 2.3 - mathematics guide. Technical report, DTU, 2003.
- N. R. Kristensen, H. Madsen, and S. B. Jørgensen. Parameter estimation in stochastic grey-box models. *Automatica*, 40(2):225 – 237, 2004. ISSN 0005-1098. doi: DOI:10.1016/j.automatica.2003.10.001.
- H. Madsen. *Time Series Analysis*. Chapman & Hall, 2007.
- H. Madsen and J. Holst. Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings*, 22(1):67–79, 1995. ISSN 03787788.
- B. Perers. Dynamic method for solar collector array testing and evaluation with standard database and simulation programs. *Solar Energy*, 50(6):517–526, 1993. ISSN 0038092x.
- B. Perers. An improved dynamic solar collector test method for determination of non-linear optical and thermal characteristics with multiple regression. *Solar Energy*, 59(4-6):163–178, 1997. ISSN 0038092x.
- Solar Keymark. Homepage and database (all tested collectors according to EN12975 in Europe), 2011. URL <http://solarkey.dk/solarkeymarkdata/qCollectorCertificates/ShowQCollectorCertificatesTable.aspx>.

146 Models of the heat dynamics of solar collectors for performance testing

P A P E R F

Short-term heat load forecasting for single family houses

Authors:

Peder Bacher¹, Henrik Madsen¹, Henrik Aalborg Nielsen², and Bengt Perers³

Submitted to:

Buildings and Energy (April 2012).

¹DTU Informatics, Richard Pedersens Plads, Building 321, DK-2800 Lyngby, Denmark

²ENFOR A/S, Lyngsø Allé 3, DK-2970 Hørsholm, Denmark (URL: www.enfor.eu)

³DTU Civil Engineering, Brovej, Building 118, DK-2800 Lyngby, Denmark

Abstract

This paper presents a method for forecasting the load for space heating in a single-family house. The forecasting model is built using data from sixteen houses in Sønderborg, Denmark, combined with local climate measurements and weather forecasts. Every hour the hourly heat load for each house the following two days is forecasted. The underlying basis of the method is physical knowledge of building heat dynamics, which, combined with statistical models, leads to a grey-box modelling approach. The forecast models are adaptive linear time-series models and the climate inputs used are: ambient temperature, global radiation, and wind speed. The adaptivity over time is achieved with a recursive least squares scheme, which is computationally very efficient. Also included is a diurnal curve for modelling patterns in the residents' behavior, for example a nightly setback. The models are optimized to fit the individual characteristics for each house, such as the level of optimal adaptivity and the thermal dynamical response of the building, which are modelled with simple transfer functions. The identification of a model, which is suitable for application to all the houses, is carried out. The results show that the one-step (one hour) ahead errors are close to white noise and that practically all correlation to the climate variables are removed with the model. Furthermore the analysis of the results shows that the main sources to forecast errors are related to: unpredictable high frequency variations in the heat load signal (predominant only for some houses), shifts in resident behavior, and uncertainty of the weather forecasts for longer horizons, especially the solar radiation is a challenge.

F.1 Introduction

The transition to an energy system based on renewables requires methods for forecasting of energy load and production. In Denmark around 40% of the total energy consumption is related to buildings and around 29% of the energy for space heating is covered by individual oil or gas fired furnaces (Danish Commission on Climate Change Policy, 2010b), which is neither an economically feasible nor environmentally friendly technology. The Danish Commission on Climate Change Policy recommends replacement with alternative technologies, especially heat pumps, since this is one of the socio-economically cheapest initiatives in the transition to an energy system without fossil fuels in Denmark. Hence, new and alternative technologies for building space heating based on renewable energy production are of high interest, both for individual and district heating. Especially electrical heating systems since large amounts of wind power are available, which creates a need for flexible load in order to absorb the

volatile production. As the level of electrical load increase, even load-shifting in shorter periods of time for peak-shaving of the diurnal electrical consumption is a valuable service to the grid (Danish Commission on Climate Change Policy, 2010a). Flexible load can be achieved with thermal energy buffering, both in individual heating and district heating, where huge thermal storage capacity is available. Several studies are considering the possibilities for flexible heating, for example Pedersen et al. (2011) and Chen (2001) who presents methods for energy storage in the thermal mass of the building, and Reddy et al. (1991) and Henze et al. (2004) who consider load-shifting for cooling of buildings. The present paper presents a method for forecasting of the power load for space heating in a single-family house. The heat load forecasts can be used as input to model predictive control, which can be used for load-shifting, for example for operation under energy markets, where relocation of load to periods with cheap energy will be rewarded. The method can just as well be used for forecasting of cooling load and used for load-shifting with cool thermal storage. Perers et al. (2011) presents solar combisystems, which is a heating system based on a solar thermal collector and electrical heating, where a hot water tank is used for thermal energy storage. Forecasting of the heat load is vital for optimal and effective use of the thermal storage in such a system.

Forecasting of the load for space heating is carried out for sixteen houses in Sønderborg, Denmark. Every hour a new forecast is calculated of the hourly heat load up to 42 hours ahead. The houses are generally built in the sixties and seventies, with a floor plan in the range of 85 to 170 m², and constructed in bricks. Climate observations - which are measured at the local district heating plant within 10 kilometers from the houses - together with numerical weather predictions (NWP) are used as input to the forecasting model. The NWP are from the HIRLAM-S05 model and provided by the Danish Meteorological Institute. For each house only the total heat load, including both space heating and water heating, is available. The heat load signal is first separated into two signals: a signal for space heating and a signal for water heating. The space heating signal is then forecasted. The splitting allows for a clear view of the effects stemming from heat loss to the ambient and heat gains from solar radiation etc., since the noisy peaks from hot water heating is filtered out. The indoor temperature is not available, which is accounted for in the models by including a diurnal curve to model nightly setback and behavioral patterns of the residents e.g. heat from electrical appliances used for cooking.

Very many approaches to load forecasting are found in the literature. A good overview of references are given by Mestekemper (2011), who built load forecasting models using dynamic factor models. Dotzauer (2002) use a model based on the ambient temperature and a weekly pattern for forecasting of the heat load in district heating, i.e. the total heat load for many houses. Zhou et al. (2008) use a grey-box model based on transfer functions for building thermal load predic-

tion and validates it on a 50 floors multi purpose building. The models applied in the present study are originally developed and used for forecasting of heat load in district heating, as described in (Nielsen et al., 2000) and (Nielsen and Madsen, 2006). The total heat load for many houses together have less high frequency variation, due to the averaging effects, compared to the heat load for a single house. Emphasis in the present study is put on the variability in heat load among the individual houses, for example some react more than others to solar radiation, and especially the diurnal pattern is very different among the studied houses.

The paper starts with a section in which the data and the NWP are described. This is followed by a presentation of the modelling approach and the model identification, where a suitable forecasting model is found together with a dynamic model for the remaining noise. Finally the results are presented, and the method is discussed and concluded.

F.2 Data

The data used in the study consist of heat load measurements from sixteen houses in Sønderborg, Denmark, and local climate measurements and NWPs. All times are in UTC and the time stamp for average values are set to the end of the time interval.

F.2.1 Heat load measurements

The houses are typical Danish single family houses from the sixties and seventies. Only houses with radiator heating is considered. A single signal for each house is used, which consist of both the energy for space heating and hot water heating. The heat load measurements consist of 10 minutes average values. Time series plots over the entire period, spanning nearly two and a half years, for four of the houses are shown in Figure F.1. Also shown, with red lines, is the distribution over time, which are estimates of the 0%, 2%, ..., 98%, 100% quantiles. They are estimated using local quantile regression (Koenker, 2005), where the weighting is local in time. They clearly indicate that the distribution of the heat load is heavily skewed, for example only two percent of the values are between the two upper lines, which cover more than half of the range. The reason for this skewness is seen from Figure F.2, where 10 days of heat load for the same four houses is plotted. The heat load for water heating consist of high frequency spikes added to the more slowly varying space heating signal. The highest peaks

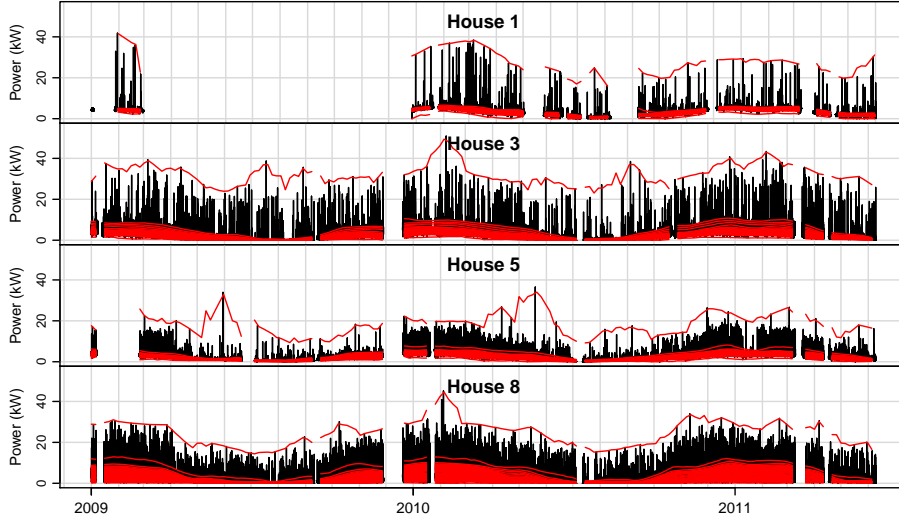


Figure F.1: The heat load for four selected houses over the entire period, which is nearly spanning two and a half years. The red lines are estimates of the 0%, 2%, ..., 98%, 100% quantiles, which indicate the distribution of the heat load at a given time.

are from showers and cause the high skewness. Since it is wanted to study the space heating part, then each signal is splitted into a space heating part and a water heating part - with the method described in (Saint-Aubain, 2011) used causally. On the figure the part of the signal identified as water heating is marked with red, note that it is added on top of the space heating signal in the plot. After the splitting the series are resampled into hourly average values. The hourly space heating for a single house is denoted by

$$\{Q_t; t = 1, \dots, N\} \quad (\text{F.1})$$

where $N = 21144$ and the unit is kW. Notice that no distinguishment in between the houses is used in the notation, but when the results are presented the house number, ranging from 1 to 16, is clearly stated.

F.2.2 Local climate observations

The local climate observations are from a weather station at the district heating plant in Sønderborg, which is less than 10 kilometers from the houses. The data

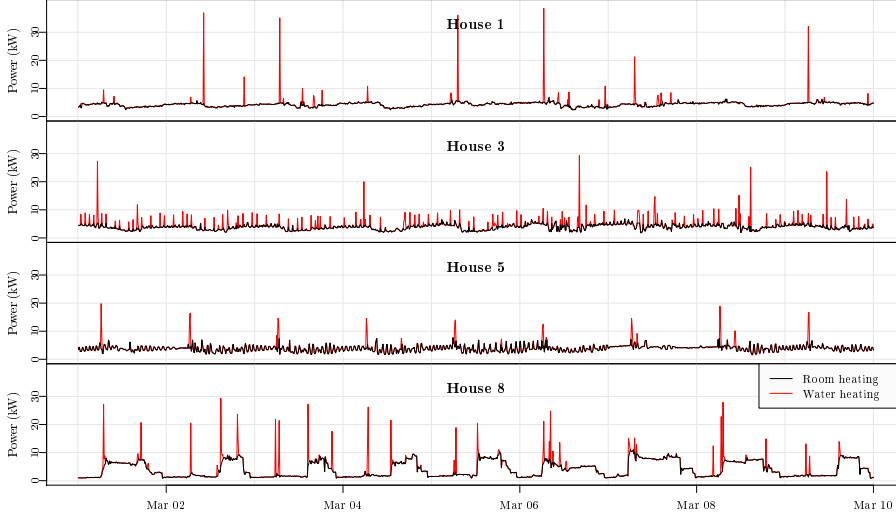


Figure F.2: Heat load for four selected houses for the first 10 days of March in 2010. The peaks marked with red are the parts which are identified as water heating and the black line is the space heating part. Note that the water heating is added on top the space heating signal.

is resampled to hourly average values and the following time series are used:

$$\text{Ambient temperature: } \{T_t^{\text{a,obs}}; t = 1, \dots, N\} \quad (\text{F.2})$$

$$\text{Global radiation: } \{G_t^{\text{obs}}; t = 1, \dots, N\}$$

$$\text{Wind speed: } \{W_t^{\text{s,obs}}; t = 1, \dots, N\}$$

F.2.3 Numerical weather predictions

The numerical weather predictions (NWP) used for the forecasting are provided by the Danish Meteorological Institute. The NWP model used is DMI-HIRLAM-S05, which has a 5 kilometer grid and 40 vertical layers (DMI, 2011). The NWP consist of time series of hourly values for climate variables, which are updated four times per day and have a 4 hour calculation delay (e.g. the forecast starting at 00:00 is available at 04:00). Since a new two day heat load forecast is calculated every hour, then - in order to use the latest available information - every hour the latest available NWP value for the k 'th horizon at

time t is picked as

$$\begin{aligned}
 \text{Ambient temperature (}^\circ\text{C)}: & T_{t+k|t}^{\text{a,nwp}} \\
 \text{Global radiation (W/m}^2\text{)}: & G_{t+k|t}^{\text{nwp}} \\
 \text{Wind speed (m/s)}: & W_{t+k|t}^{\text{s,nwp}} \\
 \text{Wind direction (}^\circ\text{azimuth)}: & W_{t+k|t}^{\text{d,nwp}}
 \end{aligned} \tag{F.3}$$

F.2.4 Combining local observations with NWP

To include the building heat dynamics in an efficient way, the inputs are low-pass filtered as explained in Section F.3.3. Hence, for the forecast calculated at time t , past values of the inputs are being used. In order to use the information embedded in the local measurements they are combined with the NWPs. The combining is carried out by forming the time series for each of the inputs at time t , for a specific horizon k , by

$$\begin{aligned}
 \{T_{t+k|t}^{\text{a}}\} &= \{\dots, T_{t-1}^{\text{a,obs}}, T_t^{\text{a,obs}}, T_{t+1|t}^{\text{a,nwp}}, T_{t+2|t}^{\text{a,nwp}}, \dots, T_{t+k|t}^{\text{a,nwp}}\} \\
 \{G_{t+k|t}\} &= \{\dots, G_{t-1}^{\text{obs}}, G_t^{\text{obs}}, G_{t+1|t}^{\text{nwp}}, G_{t+2|t}^{\text{nwp}}, \dots, G_{t+k|t}^{\text{nwp}}\} \\
 \{W_{t+k|t}^{\text{s}}\} &= \{\dots, W_{t-1}^{\text{s,obs}}, W_t^{\text{s,obs}}, W_{t+1|t}^{\text{s,nwp}}, W_{t+2|t}^{\text{s,nwp}}, \dots, W_{t+k|t}^{\text{s,nwp}}\} \\
 \{W_{t+k|t}^{\text{d}}\} &= \{\dots, W_{t-1|t}^{\text{d,nwp}}, W_{t|t}^{\text{d,nwp}}, W_{t+1|t}^{\text{d,nwp}}, W_{t+2|t}^{\text{d,nwp}}, \dots, W_{t+k|t}^{\text{d,nwp}}\}
 \end{aligned} \tag{F.4}$$

Notice that local observations are not available for the wind direction.

F.3 Models

As mentioned earlier the applied models are similar to the models used by Nielsen and Madsen (2006) for modelling of the total heat load for many houses. The models are based on prior physical knowledge of the heat dynamics of buildings, which in combination with statistical time series models, forms a grey-box modelling approach. This allows for inclusion of heat transfer effects related to the climate variables in a combination with time adaptivity for modelling of changing condition. Furthermore, in order to describe of patterns in resident behavior, a diurnal curve is included. The forecasting models are fitted, by optimizing the parameters to minimize the RMSE in an off-line setting. The fitting is carried out separately for each house and for each horizon k , which means that the same model formulation - i.e. same inputs and model structure - is used, but the parameter values for each house and horizon can be different.

F.3.1 Time adaptive models

The models are fitted with the k -step recursive least squares scheme described by Bacher et al. (2009). This means that the coefficients in the model can change over time and thereby adapt optimally to changing conditions. The coefficients are recursively updated, which means that only a few matrix operations are required to make a new forecast, hence the scheme is computationally very fast. It is a recursive implementation of a weighted least squares estimation, where the weights are exponentially decaying over time. A single parameter is required, the forgetting factor λ , which determines how fast input data is down-weighted. The weights are equal to

$$w(\Delta t) = \lambda^{\Delta t} \quad (\text{F.5})$$

where Δt is the age of the data in hours. This means that for $\lambda = 0.98$ the weights are halved in 34 hours, for $\lambda = 0.995$ they are halved in 138 hours (~ 6 days), and for $\lambda = 0.999$ in 693 hours (~ 29 days).

F.3.2 Diurnal curve

A curve for describing systematic diurnal patterns in the heat load is included in the models, which for example can be caused by a nightly setback and free heat from electrical appliances. The curve is modelled as a harmonic function using a Fourier series

$$\mu(t_{\text{tod}}, \alpha_{\text{diu}}) = \sum_{i=1}^{n_{\text{har}}} \alpha_{i,1}^{\text{diu}} \sin\left(\frac{t_{\text{tod}} i \pi}{12}\right) + \alpha_{i,2}^{\text{diu}} \cos\left(\frac{t_{\text{tod}} i \pi}{12}\right) \quad (\text{F.6})$$

where t_{tod} is the time of day in hours at time t and n_{har} is the number of harmonics included in the Fourier series. For all the applied models a curve is fitted for working days and another curve for weekends.

F.3.3 Low-pass filtering for modelling of building dynamics

The heat dynamics of a building can be described by lumped parameter RC-models, see for example (Madsen and Holst, 1995), (Braun and Chaturvedi, 2002), and (Jiménez et al., 2008). As described by Nielsen and Madsen (2006) the response in the heat load to changes in the climate variables can be modelled with rational transfer functions, which is a description of the low-pass filtering

effect of the building with an RC-model. In the present models the simplest first order low-pass filter, with a stationary gain equal to one, is used. This is a model of the building heat dynamics formed by an RC-model with a single resistance and a single capacitor. As an example the transfer function from the ambient temperature to the heat load is

$$Q_t = H_a(q)T_t^a \quad (\text{F.7})$$

where

$$H_a(q) = \frac{1 - a_{T_a}}{1 - a_{T_a}q^{-1}} \quad (\text{F.8})$$

and where q^{-1} is the backward shift operator ($q^{-1}x_t = x_{t-1}$) (see (Madsen, 2007)) and $a_{T_a} \in [0, 1]$ is a parameter, which is equivalent to the time constant for the part of the building affected by changes in ambient temperature. A building with a high thermal mass and good insulation will have a relatively high a_{T_a} , hence the filter parameter needs to be estimated for each building in order to describe the heat dynamics properly. First order low-pass filters are also applied for wind speed and global radiation, with the filter parameter is estimated for matching of the response of the building to each of the climate variable separately.

F.3.4 Parameter optimization

As described above several parameters needs to be optimized for each house and horizon. The optimization is carried out in an off-line setting by minimizing the root mean square error for each of the sixteen houses and for each horizon $k = 1, \dots, 42$ separately. This does require some computational power, especially the low-pass filtering of the inputs. Therefore a simple bisectioning scheme is applied for the optimization, since this allows for performing a filtering of the inputs only once for parameter values in a given range. Then these series can be used for optimization for all the houses and horizons.

The following parameters are optimized

- The forgetting coefficient: λ
- The number of harmonics in the diurnal curve: n_{har}
- The coefficients for input low-pass filters: a_{T_a} , a_G , and a_{W_s}

F.4 Model identification

Forecasting models, which include different types of heat transfer effects related to the climate variables, are applied to identify which of the inputs are important to include. Furthermore different ways for the inputs to enter the model are tried. See (Nielsen and Madsen, 2006) for a description of how a physical model can be rewritten into the identifiable models, which are used here. The model which include all energy contributions is

$$Q_{t+k} = \hat{Q}_{t+k|t} + e_{t+k} \quad (\text{F.9})$$

where

$$\hat{Q}_{t+k|t} = Q_a + Q_g + Q_w \quad (\text{F.10})$$

where the Q_{name} variables on the right side of the equation represents the heat contributions from the considered heat transfers, which are described below.

F.4.1 First step in model selection

To select a suitable forecasting model a forward selection approach is used. In the first step the *model_D*

$$\hat{Q}_{t+k|t} = \alpha_{ia} + \mu(t_{tod}, \alpha_{diu}) \quad (\text{F.11})$$

which do not include the climate inputs, is fitted. In this model the heat load is simply modelled as a diurnal curve with an offset. Note that α_{ia} then represents a constant difference between the indoor and the ambient temperature and the diurnal part will try to capture diurnal patterns in both the indoor and ambient temperature. Due to the time adaptive scheme the model will be able to track the slow changes in the temperatures over the year. Finally note also that the coefficients could have been denoted with: a t to indicate that they are changing over time, a house number to indicate that they are fitted to each house, and a k to indicate that the model is fitted for each horizon separately, but this have been left out for better readability.

To find out if there is useful information available in the climate series as inputs to the model, the cross-correlation function (CCF), see (Box et al., 1976), between the one-step ahead ($k = 1$ hour) error for *model_D* and the available input series is calculated - which is preferred over the NWP, as it is mainly the observations which are used for the one-step ahead forecast. Since it is too

cumbersome to analyze the error series for each house separately, the average error

$$\bar{e}_{t+k} = \frac{1}{n_{\text{houses}}} \sum_{i=1}^{n_{\text{houses}}} e_{t+k}^{\text{House } i} \quad (\text{F.12})$$

where $n_{\text{houses}} = 16$ are the number of houses and $e_{t+k}^{\text{House } i}$ is the error series for House i , is used. In this way the CCF to the inputs is summarized for all the houses in a single plot. This will of course only show if an input is generally important to include and not the effects for each individual house. The effects related to each house - which are different - are considered in later parts of the paper.

The CCF between the average errors from $model_D$ to the inputs can be seen in Figure F.3a. Clearly very significant correlations between the error and both the ambient temperature and the global radiation, but apparently none to the wind speed, are found. It is decided to add the ambient temperature as input to the model, which leads to $model_A$

$$\hat{Q}_{t+k|t} = Q_a \quad (\text{F.13})$$

where

$$Q_a = \alpha_i + \mu(t_{\text{tod}}, \alpha_{\text{diu}}) + \alpha_a H_a(q) T_{t+k|t}^a \quad (\text{F.14})$$

The $H_a(q)$ is the low-pass filter modelling the dynamics of the building envelope, i.e. the response of in heat load to changes in ambient temperature. Notice that the intercept α_i is representing a constant indoor temperature modified by the diurnal curve.

In order to find out if any gain in forecasting performance is achieved from $model_D$ to $model_A$ the root mean square error is used for evaluation as described in the following.

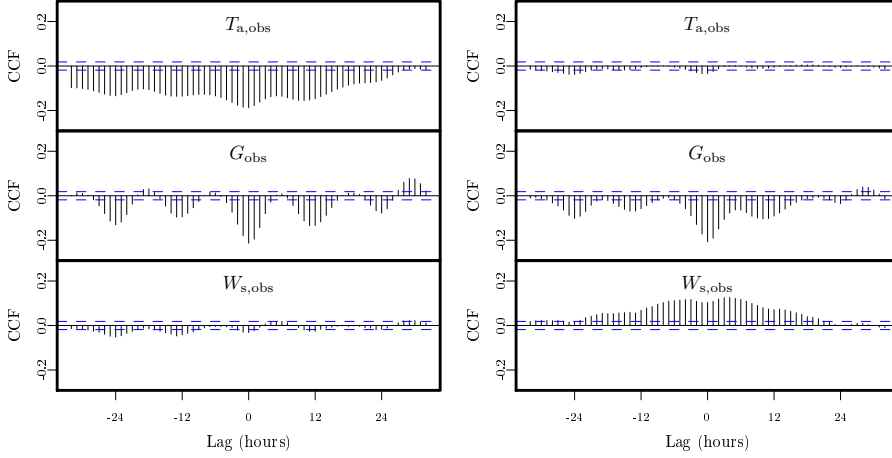
F.4.1.1 Root mean square error evaluation

To evaluate the models the root mean square error (RMSE) for the k 'th horizon

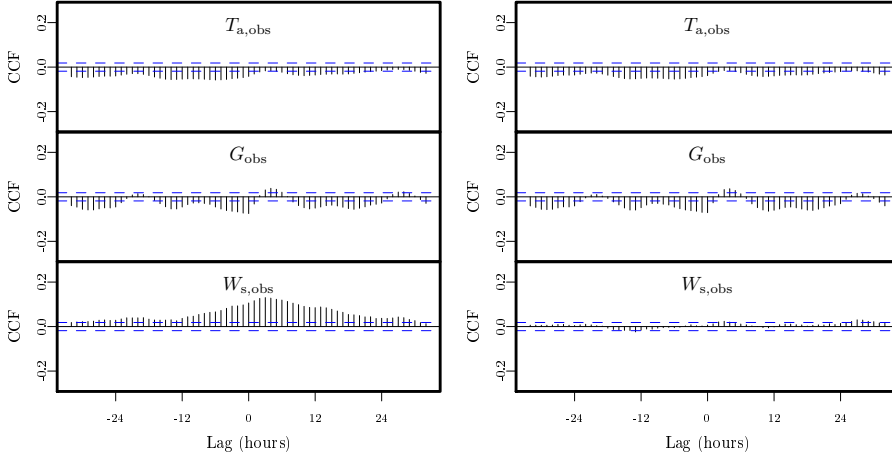
$$\text{RMSE}_k = \left(\frac{1}{N} \sum_{t=1}^N e_{t+k}^2 \right)^{\frac{1}{2}} \quad (\text{F.15})$$

is used together with the RMSE_k improvement

$$I_{\text{RMSE}} = 100 \cdot \frac{\text{RMSE}_{\text{ref}} - \text{RMSE}}{\text{RMSE}_{\text{ref}}} (\%) \quad (\text{F.16})$$



(a) Step one: CCF for $model_D$ errors to the inputs. (b) Step two: CCF for $model_A$ errors to the inputs.



(c) Step three: CCF for $model_{A,G}$ errors to the inputs. (d) Step four: CCF for $model_{A,G,W}$ errors to the inputs.

Figure F.3: The cross-correlation function (CCF) between the average error series for the one-step ahead forecast (horizon $k = 1$) and the local observations of the inputs, since the inputs for the one-step ahead forecasts are primarily formed by the observations. The plots are for the errors from the selected model in the four steps of model identification carried out.

over the currently selected model as reference. It is noted values not in the heating season - which starts the 15'th of September and ends 15'th of May in Denmark - and values before the 15'th of March 2009, which is used as a burn-in period, are excluded from the RMSE_k calculation. For evaluation of the inclusion of ambient temperature, the RMSE_k improvement for model_A over model_D is calculated. The average improvement for all horizons (from $k = 1$ to $k = 42$) for each house is plotted in Figure F.4a. A RMSE_k improvement for the each house in the range from 5 to 25 percent and around 14 percent in average is achieved. This is clearly a significant improvement, hence model_A is preferred over model_D .

F.4.2 Second step in model selection

To explore the possibilities for further expansion of model_A the CCFs from the average errors (defined in Equation (F.12)) to each of the climate series are calculated. They are plotted in Figure F.3b. The correlation to the ambient temperature is much lower than for model_D and the correlation to the global radiation is more or less the same. The correlation to the wind speed has increased, most likely this correlation was "overshadowed" by the correlation to the ambient temperature for the model_D errors. Notice that there is a significant correlation decaying over 12 to 24 hours to the lagged inputs, which indicates that dynamics should be included by low-pass filtering.

To find the most important extension of model_A several extensions involving the global radiation or the wind speed are fitted (i.e. the RMSE_k is minimized by tuning the parameters listed in Section F.3.4 for each house).

The first considered expansion is $\text{model}_{A,G}$

$$\hat{Q}_{t+k|t} = Q_a + \alpha_g H_g(q) G_{t+k|t} \quad (\text{F.17})$$

where the heat gain from solar radiation is included by letting the global radiation enter through a low-pass filter, which models the dynamic response from the global radiation to the heat load of the building. The second expansion is $\text{model}_{A,W}$

$$\hat{Q}_{t+k|t} = Q_a + \alpha_{ws} H_w(q) W_{t+k|t}^s \quad (\text{F.18})$$

where the cooling of the building from wind is modelled by letting the wind speed enter through a low-pass filter. This is a model of wind cooling not depending on the ambient temperature, however - due to the time-adaptive modelling scheme - it does include the slow changes in temperature difference between indoor and

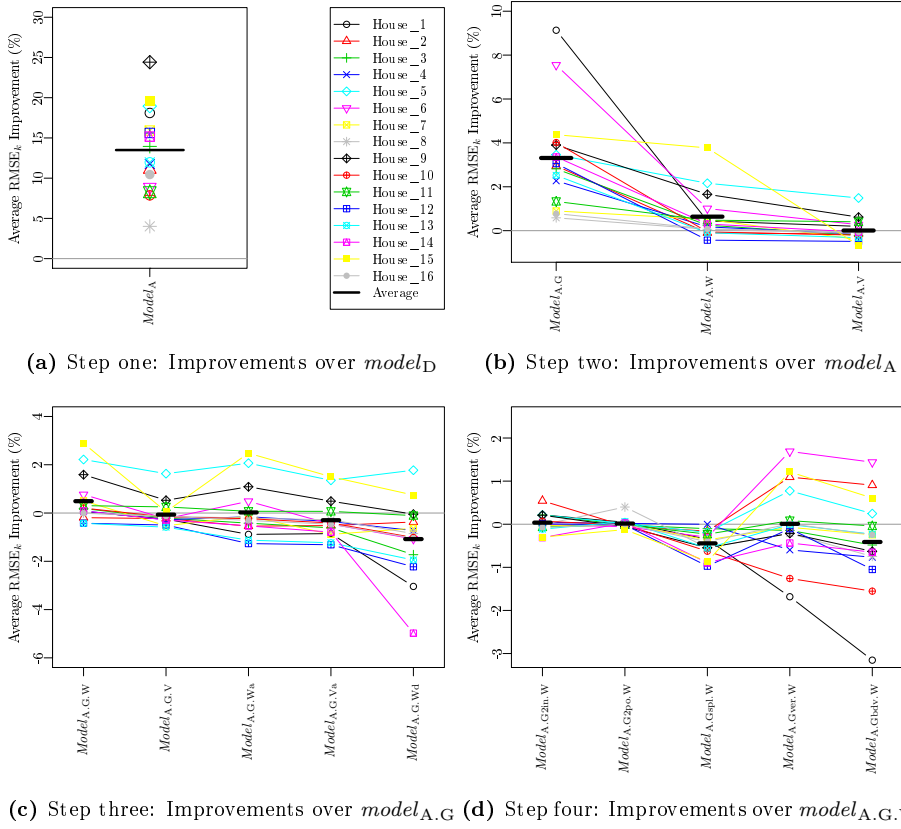


Figure F.4: Improvements over the previously selected model for the models in each step of the selection.

ambient temperature. The third expansion is $model_{A,V}$

$$\hat{Q}_{t+k|t} = Q_a + \alpha_{\text{vent}} W_{t+k|t}^s \quad (\text{F.19})$$

which include an effect of ventilation modelled by inputting the instant effect of wind speed to the heat load.

The $RMSE_k$ improvements averaged over all horizons for each house for the considered expansions is plotted in Figure F.4b. It is seen that the performance increase is highest for all the houses for $model_{A,G}$, hence this model is preferred and used for expansion in the following step. In the remaining of the paper the heat contribution from solar radiation is denoted by

$$Q_g = \alpha_g H_g(q) G_{t+k|t} \quad (\text{F.20})$$

F.4.3 Step three: Inclusion of wind speed in the model

In the third step of the model identification several ways of including the wind speed is considered. First the CCFs between the one-step ahead error, from the model selected in the previous step, $model_{A,G}$ and the inputs, are studied to see if any useful information is remaining in the climate series.

First it is noted that the correlation to the global radiation has decreased compared to the CCF plot for $model_A$, but that some is still left, indicating that the dynamic effects are not entirely described by the model. Secondly it is noted that there is a significant cross-correlation to the wind speed and therefore it is chosen to seek an expansion which include the wind speed. Five different ways of letting the wind speed enter the model are tried as described in the following.

The first expansion is formed by adding the wind speed through a low-pass filter for modelling of cooling of the building in $model_{A,G,W}$

$$\hat{Q}_{t+k|t} = Q_a + Q_g + \alpha_{ws} H_w(q) W_{t+k|t}^s \quad (\text{F.21})$$

and, for modelling ventilation, the instant effect of wind speed is added in $model_{A,G,V}$

$$\hat{Q}_{t+k|t} = Q_a + Q_g + \alpha_{ws} W_{t+k|t}^s \quad (\text{F.22})$$

In the two models above the wind speed enter the model without the interaction with ambient temperature, which means that the temperature difference between the indoor and ambient temperature is modelled as constant and that

changes are only tracked with the adaptivity of the model. In the following two expansions the interaction is also included, with a filter in $model_{A.G.Wa}$

$$\hat{Q}_{t+k|t} = Q_a + Q_g + \alpha_{ws} H_w(q) W_{t+k|t}^s + \alpha_{wsa} H_w(q) W_{t+k|t}^s T_{t+k|t}^a \quad (F.23)$$

and as an instant effect in $model_{A.G.Va}$

$$\hat{Q}_{t+k|t} = Q_a + Q_g + \alpha_{ws} W_{t+k|t}^s + \alpha_{wsa} W_{t+k|t}^s T_{t+k|t}^a \quad (F.24)$$

Finally the wind speed input coefficient is conditioned on the wind direction in $model_{A.G.Wd}$

$$\hat{Q}_{t+k|t} = Q_a + Q_g + \sum_{i=1}^4 \alpha_i H_w(q) K(u) W_{t+k|t}^s T_{t+k|t}^a \quad (F.25)$$

where the kernel function

$$K(u) = (1 - |u|) \mathbf{1}_{\{|u| \leq 1\}} \quad (F.26)$$

with

$$u = \left((W_{t+k|t}^{d,nwp} + 45 + (i-1) \cdot 90) \bmod 4 \right) - 1 \quad (F.27)$$

makes four input series, which are linearly interpolated as a function of the wind direction. The center of the kernels is thus at the most prevailing wind directions in Denmark, especially southwest in the winter period (Cappelen and Jørgensen, 1999).

The plot in Figure F.4c shows the improvements over $model_{A.G}$ for the five models. The improvement is quite different for each house, for some it is negative, which is because the forecasting model becomes over-parametrized. It is also seen that the pattern of the improvement among the houses are quite similar for the five models, which indicates that for some houses the wind have a more prevalent effect than for others. Since $model_{A.G.W}$ generally have the most positive improvement and since it is the simplest extension, it is preferred over the others. In the remaining of the paper the model part describing the effect of wind is denoted with

$$Q_w = \alpha_{ws} H_w(q) W_{t+k|t}^s \quad (F.28)$$

F.4.4 Step four: Enhancement of the solar model part

In the final step the model part for solar radiation is enhanced in different ways, as described in the following. Studying the CCFs for $model_{A.G.W}$ in Figure F.3d

only very little correlation between the errors and the inputs are seen. There is a little left to the ambient temperature and solar radiation. From studying plots of the forecasts it is found that it could be possible to improve the model part, where the effect of solar radiation is included.

First an additional input for the solar gain is added the model, this is $model_{A.G2in.W}$

$$\hat{Q}_{t+k|t} = Q_a + Q_g + \alpha_{g2} H_{g2}(q) G_{t+k|t} + Q_w \quad (F.29)$$

This will allow for an additional dynamic response of the building to solar radiation. Notice that an additional filter coefficient for the $H_{g2}(q)$ filter is fitted here.

Secondly the solar radiation part is enhanced by using a two-pole filter instead of the one-pole filter

$$\hat{Q}_{t+k|t} = Q_a + \alpha_g H_{2pol}(q) G_{t+k|t} + Q_w \quad (F.30)$$

where

$$H_{2pol}(q) = \frac{1 - a_1 - a_2}{1 - a_1 q^{-1} - a_2 q^{-2}} \quad (F.31)$$

and

$$a_1 = a_{g1} + a_{g2} \quad \text{and} \quad a_2 = -a_{g2} \quad (F.32)$$

The two filter coefficients thereby relate to different dynamics: a_{g1} is related to the highest time constant (slow response) and a_{g2} is related to the lowest time constant of the building (fast response).

In the third model the solar radiation is separated into three inputs: one for the morning, one for the noon, and one for the evening. This allows for the building to have different solar gains during the day. The $model_{A.Gspl.W}$ is

$$\begin{aligned} \hat{Q}_{t+k|t} = & Q_a + \alpha_{g1} H_g(q) G_{t+k|t}^{\text{morning}} + \alpha_{g2} H_g(q) G_{t+k|t}^{\text{noon}} \\ & + \alpha_{g3} H_g(q) G_{t+k|t}^{\text{evening}} + Q_w \end{aligned} \quad (F.33)$$

The three inputs are

$$\begin{aligned} G_{t+k|t}^{\text{morning}} &= G_{t+k|t} \mathbf{1}_{\{t+k \in [t_{\text{rise}}, t_{\text{rise}} + \Delta t]\}} \\ G_{t+k|t}^{\text{noon}} &= G_{t+k|t} \mathbf{1}_{\{t+k \in [t_{\text{rise}} + \Delta t, t_{\text{set}} - \Delta t]\}} \\ G_{t+k|t}^{\text{evening}} &= G_{t+k|t} \mathbf{1}_{\{t+k \in [t_{\text{set}} - \Delta t, t_{\text{set}}]\}} \end{aligned} \quad (F.34)$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function, t_{rise} and t_{set} is the time of sunrise and sunset, respectively, and $\Delta t = (t_{\text{set}} - t_{\text{rise}})/3$ is a third of the day length.

Finally two more possible enhancement is applied by projecting the solar radiation onto a vertical surface tracking the solar azimuth angle. The projection to vertical is carried out by first splitting the global radiation into a direct and a diffuse component as in (Ruiz-Arias et al., 2010) and onto a vertical surface with the Hay and Davies model (Hay and Davies, 1980), see also (Reindl et al., 1990). The effect of the projections is mostly an increased level of solar radiation in the morning and in the evening (or afternoon), when the sun elevation is low. In *model*_{A.Gver.W}

$$\hat{Q}_{t+k|t} = Q_a + \alpha_g H_g(q) G_{t+k|t}^{\text{ver,tr}} + Q_w \quad (\text{F.35})$$

the total vertical radiation is inputted and in *model*_{A.Gbdv.W}

$$\hat{Q}_{t+k|t} = Q_a + \alpha_{g1} H_g(q) G_{t+k|t}^{\text{beam,tr}} + \alpha_{g2} H_g(q) G_{t+k|t}^{\text{diffuse,tr}} + Q_w \quad (\text{F.36})$$

the direct (or beam) and the diffuse component is inputted separately.

The models are fitted to each house and the improvements over *model*_{A.G.W} are calculated and plotted in Figure F.4d. For *model*_{A.G2in.W} and *model*_{A.G2po.W} only a little difference in performance is seen, and for *model*_{A.Gspl.W} the performance has decreased. These three models become over-parametrized, however it is noted that for short horizons the improvement for *model*_{A.Gin2.W} is positive for all the houses. For the last two: *model*_{A.Gver.W} and *model*_{A.Gbdv.W}, the change in performance depends on the house. One interesting pattern is that the houses with the highest change (both decreased and increased) are the houses which benefit most from addition of the solar radiation to the model, as seen in the plot in Figure F.4b. This indicates that increased performance can be gained with a model which modifies the solar gain over the day. Such a model should learn an optimal diurnal solar gain curve for each house. In general no significant overall increase in performance is found for any of the five models, hence the model selection is ended and the results in the remaining of the paper are from *model*_{A.G.W}, together with a model of the noise, which is described in the following section.

F.5 Noise model

Considering the auto-correlation for the one-step ahead error for the houses, shown in the upper plot of Figure F.5, it is found that a model is useful for describing dynamical information embedded in the errors. A simple auto-regressive (AR) model is fitted to the errors with the recursive least squares scheme (Bacher et al., 2009). The AR(1) model

$$e_{t+k} = a_e e_t + e_{t+k}^{\text{noise}} \quad (\text{F.37})$$

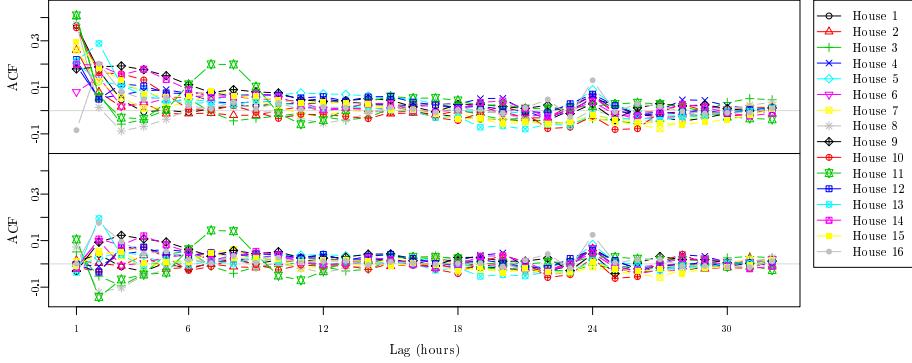


Figure F.5: The auto-correlation function (ACF) for each house. The upper plot is the ACF of the errors before the AR(1) noise model is applied and the lower plot shows the ACF of the errors after.

is fitted for the errors from the selected model $model_{A.G.W}$ for each horizon k . The ACF of the noise error e_{t+k}^{noise} can be seen in the lower plot of Figure F.5. Compared to the upper plot almost all of the auto-correlation for lag 1 is removed. Clearly, some of the houses still have significant auto-correlation left and for the short horizons an error model, which include more lags could improve the performance. However it was tried to include one more lag (lag 2), but this did only improve the overall performance marginally, mainly because no performance improvement is achieved on longer horizons. The houses which have the highest ACF (in particular House 11 and 16) have some high frequency oscillations embedded in the heat load signal, as described in the following section where the results are discussed. The average $RMSE_k$ improvement over all horizons is in the range of 0.35% to 6.7%, hence a quite significant improvement, especially for some of the houses. The $RMSE_k$ improvement for the one-step ahead forecasts is in the range 1.3% to 19%, which clearly shows that the noise model is most important for short horizons.

F.6 Results

In this section the results from forecasting with the selected model are presented and discussed. First the parameters, which are fitted for each house, are reported and then the performance for individual houses is discussed.

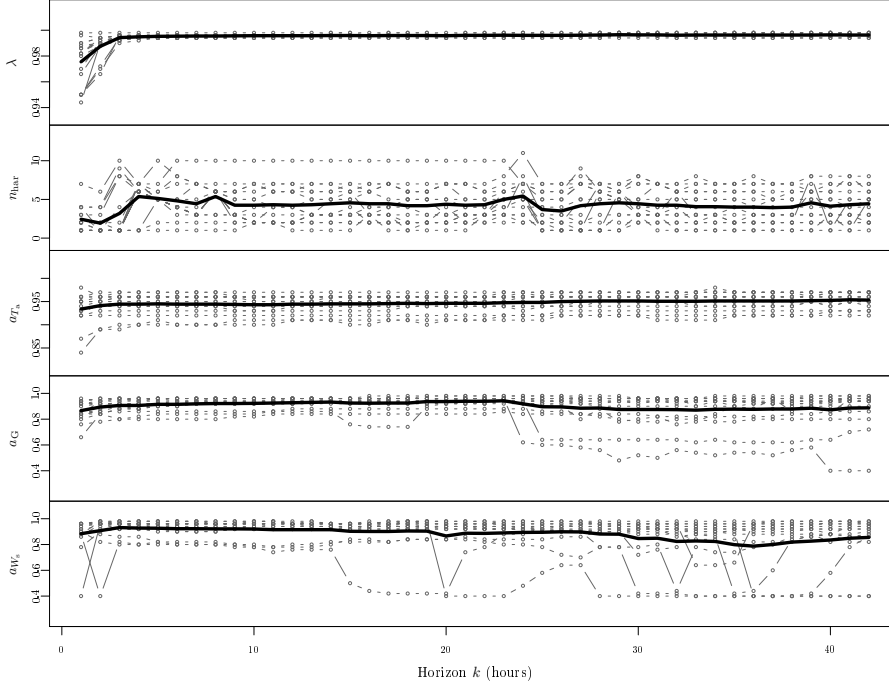


Figure F.6: Values of the parameters fitted for each house.

F.6.1 Model parameters

The parameters which are fitted for each house are listed in Section F.3.4. Since there is a value for each horizon for each house for each parameter, they are reported with the plots in Figure F.6. It is the general patterns which are discussed in the following. Starting with the upper most plot in the figure, which is of the forgetting coefficient λ in the recursive least squares scheme, it can be seen that it has a tendency to be lower for the first couple of horizons: for $k = 1$ the average over all the house is 0.9755, which means that the weighting of the input data is halved in only 28 hours. This fast forgetting is most likely optimal, because it is profitable for the forecasting model to be able to react fast to changes in to the system, e.g. residents increase the indoor temperature or open the windows, which can be tracked on short horizons. On longer horizons the forgetting is on a stable level: for $k = 5$ the average is 0.9953 increasing to 0.9963 for $k = 42$, which means that the weighting of the input data is halved in around 8 days.

The second plot from the top in Figure F.6 is of the optimized number of

House	1	2	3	4	5	6	7	8
a_{T_a}	0.96	0.91	0.91	0.96	0.94	0.96	0.95	0.96
τ_{T_a} (hours)	27	11	10	25	16	27	21	26
House	9	10	11	12	13	14	15	16
a_{T_a}	0.94	0.97	0.95	0.96	0.95	0.94	0.96	0.92
τ_{T_a} (hours)	17	32	18	24	19	15	25	13

Table F.1: Values of optimized low-pass filter coefficient for the response from ambient temperature to heat load and corresponding RC time constant τ_{T_a} in hours.

harmonics in the diurnal curve: a higher number means that it is profitable to include harmonics up to higher frequencies in the diurnal curve. Clearly, a huge variation among the houses is found, which is very reasonable, since the diurnal patterns are very different as shown later in the paper.

The middle plot of the figure is of the optimized coefficient for the low-pass filter transfer function from the ambient temperature to the heat load. Except for the two lowest lines the variation for each house as a function of the horizon k is actually quite little (in the range of ± 0.01), which leads to the conclusion that this does describe very well, how fast the response of heat load is to changes in ambient temperature for the house. In Table F.1 the average coefficient for each house is listed together with the equivalent RC time constants. The values are within a reasonable range compared to values found in other studies (Nielsen and Madsen, 2006), (Reddy et al., 1991).

The fitted values of low-pass filter coefficient for global radiation a_G and for wind speed a_{W_s} are shown in the lower two plots of Figure F.6. The values are all in the same range, generally between 0.8 and up to near 1, but with some lower values for a couple of the houses, which are houses where the solar radiation and wind speed are not very important inputs.

F.6.2 Forecasting performance

In this section the forecasting performance is presented and discussed, especially the differences in performance among the houses. For evaluation of the performance the normalized root mean square error for each horizon

$$\text{NRMSE}_k = \frac{\text{RMSE}_k}{\bar{Q}_t} \quad (\text{F.38})$$

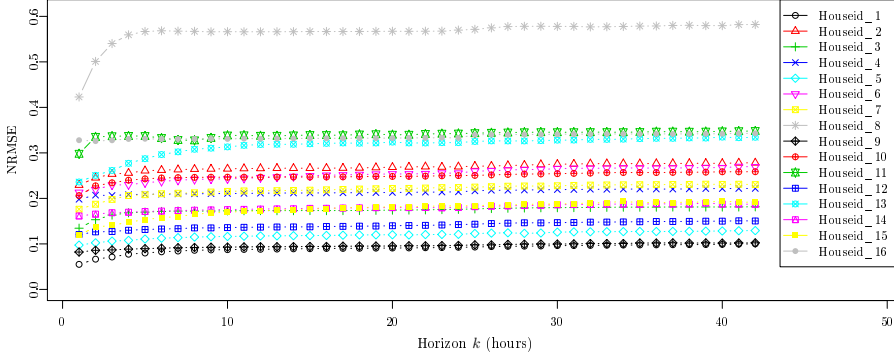


Figure F.7: The NRMSE_k as a function of the horizon k for $\text{model}_{\text{A.G.W}}$ for each house.

is used, where \bar{Q}_t is the average heat load for the house, which is calculated with the same values as used for calculation of the RMSE_k (see the text below Equation (F.15)).

The plot in Figure F.7 shows the NRMSE_k as a function of the horizon k for each house using the selected $\text{model}_{\text{A.G.W}}$. Clearly the poorest forecasting performance is for House 8. The explanation is found by considering the plot for House 8 in Figure F.8, which shows the heat load together with the 1 hour and 24 hour forecasts. The main reason for the poor forecasts is a very irregular diurnal curve. A nightly setback makes a huge difference in heat load from day to night and furthermore the time of day, where the heat is turned high again, varies among the days and is probably controlled manually by the residents. This is opposed to the nightly setback for House 10, which have a much more regular pattern that can be much better forecasted.

Another source for high errors is seen in the plots for House 2 and 16, where noisy fluctuations occur on the higher frequencies in the signals. The smaller fluctuations are probably partly from water heating, which was not well separated from the space heating, but clearly higher peaks not related to water heating are seen. For House 11 a more steady, but still quite unpredictable, fluctuation is seen, which is likely to come from some oscillation in the thermostatic control of the heat system.

The heat load signals for House 1, 9, and 15 are much less volatile. These houses are also the ones with a lower NRMSE_k , as seen from the plot in Figure F.7. The most obvious thing to notice is the deviation between the 1 and the 24 hour forecasts. Starting with the drop on the 21'st of February, which is followed well by the 1 hour forecast, but not by the 24 hour forecast. This drop is clearly

caused by solar radiation. It is a clear-sky day as seen by the high level of observed global radiation (the second uppermost plot of Figure F.8), which is also predicted by the 24 hour NWP. However the drop is not followed by the 24 hour load forecast, since the previous day was also forecasted as a clear-sky day by the 24 hour NWP, but it was not a clear-sky day as seen by the low observed level. Hence there is a much higher uncertainty on the global radiation input to the 24 hour model compared to the 1 hour model, which use mostly observations as input, and therefore the global radiation is not given much weight in the 24 hour forecasting model. From the 1st of March a sunny period begins and it can be seen how the 24 hour forecasts starts to track the mid-day drops in heat load, as more weight is put on the global radiation input due to the time adaptivity of the modelling scheme. Finally it is noted that the drop the 23rd of February and the peak the 27th of February in the heat load for House 15 are not seen in the other heat loads. It is attributed to residents behavior, which cannot be forecasted - though it is tracked in the 1 hour forecast, but with a delay.

F.7 Discussion

For the data used in the study the indoor temperature was not available and it is therefore modelled as constant modified with a diurnal curve. If the indoor temperature is available it will allow for advancements of the method by including it as an input to the model. The estimation of a time constant for the building can be carried out with higher accuracy, which will allow the method to be used for smart grid applications, such as load-shifting by use of the building structure for thermal storage.

Regarding the model part in which the solar radiation is entering, further advancements could be tried. For example more information about the individual buildings such as the azimuth angle of the walls, would provide a more detailed projection of the radiation from horizontal to vertical surfaces. However it will be favorable for operation if no specific information about the buildings is needed. The possibilities are then to try non-linear functions, which could be applied with piecewise linear or regime switching functions depending on the level of solar radiation, or a non-parametric approach could also taken with an off-line method for learning how the building respond to solar radiation over the day. Finally it is found that performance can be increased by using different models depending on the horizon, especially it is more relevant to increase the model complexity for shorter horizons.

F.8 Conclusion

A method for forecasting the load for space heating in a single-family house is presented. It is formed by adaptive linear time-series modelling techniques, using local observations and weather forecasts as input. Based on measurements from sixteen houses a model, which is suitable for all the houses, is identified by using a forward selection approach. It is shown how the forecasting performance increases when the ambient temperature, global radiation, and wind speed are added as inputs to the model. For inclusion of the heat dynamics of the building into the model simple transfer functions are used. Several further advancements to the model are tried, but this makes the model over-parametrized and results in decreased performance. In a second step an auto-regressive model is applied for modelling of the remaining dynamic information in the error. After this it is shown that almost no auto-correlation is left in the errors and thereby that the heat load is modelled very well. The model parameters, which have been fitted individually for each house, are analyzed and it is found that they provide reliable information about the dynamic response of the building. The forecasting results are then analyzed thoroughly to give insight into the sources of error, for example unpredictable behavior of the residents and uncertainty in the inputs, especially from the solar radiation weather forecasts. Finally, a discussion is given with ideas for further advancements.

Acknowledgement

Acknowledgements are given to the The Danish Council for Strategic Research, which have provided the financial support for the project "Solar/electric heating systems in the future energy system" (2104-07-0021) under which the work was carried out. The heat load and climate data was very kindly provided by Sønderborg Fjernvarme and the Danish Meteorological Institute is thanked for making their numerical weather predictions available.

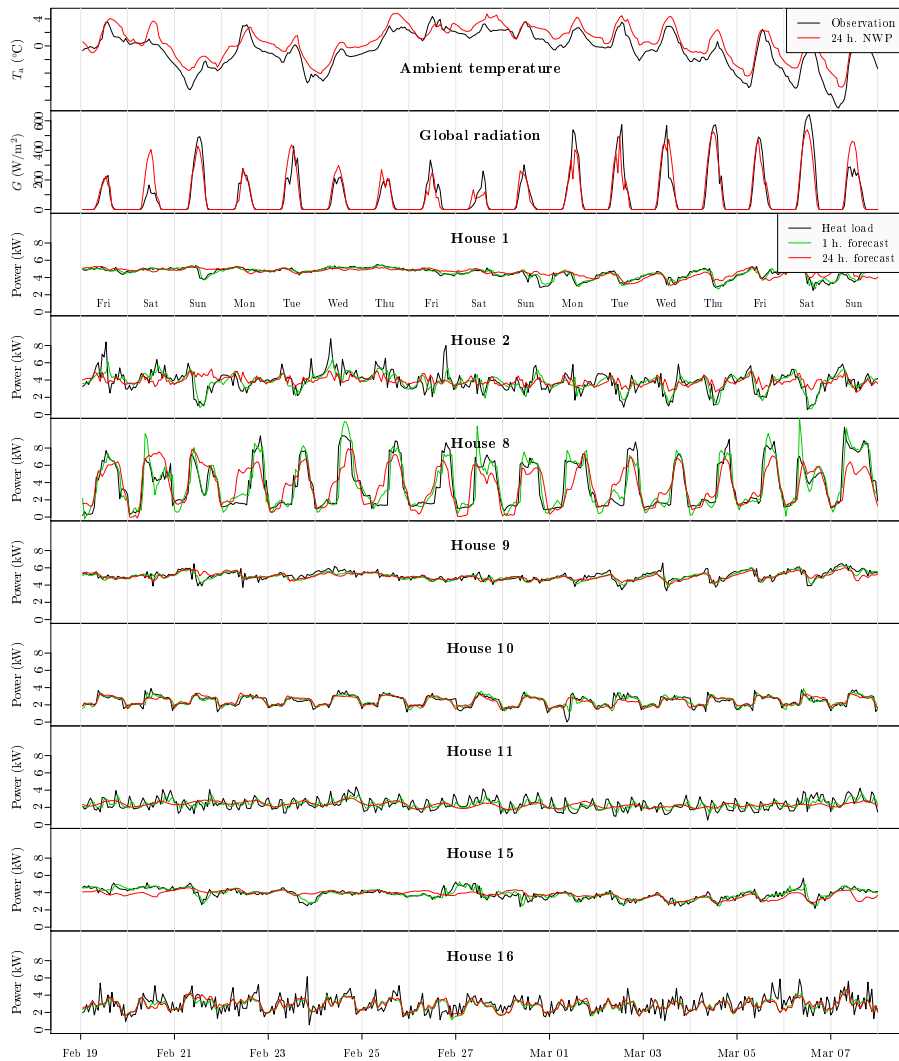


Figure F.8: Plots of inputs and forecasts for sixteen days starting at the 19th February 2010. The upper two plots are observations and the 24 hour NWP forecasts of the two most significant inputs: the ambient temperature and the global radiation. The plots below are for eight the houses of: observed heat load, 1 hour forecasts, and 24 hour forecasts.

References

- P. Bacher, H. Madsen, and H. A. Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83(10):1772–1783, 2009. ISSN 0038092x.
- G. Box, G. Jenkins, and G. Reinsel. *Time series analysis*. Holden-day San Francisco, 1976.
- J. E. Braun and N. Chaturvedi. An inverse gray-box model for transient building load prediction. *HVAC&R Research*, 8(1):73–99, 2002. doi: 10.1080/10789669.2002.10391290.
- J. Cappelen and B. Jørgensen. Observed wind speed and direction in denmark - with climatological standards normals, 1961-90. Technical report, Danish Meteorological Institute, 1999.
- T. Chen. Real-time predictive supervisory operation of building thermal systems with thermal mass. *ENERGY AND BUILDINGS*, 33(2):141–150, JAN 2001. ISSN 0378-7788.
- Danish Commission on Climate Change Policy. Green energy – the road to a danish energy system without fossil fuels, September 2010a.
- Danish Commission on Climate Change Policy. Grøn energi - vejen mod et dansk energisystem uden fossile brændsler. Technical report, Danish Energy Agency, September 2010b. Dokumentationsdelen til Klimakommissionens samlede rapport.
- DMI. Danish Meteorological Institute, DMI-HIRLAM-S05, 2011. URL http://www.dmi.dk/eng/index/research_and_development/dmi-hirham-2009.htm.
- E. Dotzauer. Simple model for prediction of loads in district-heating systems. *Applied Energy*, 73(3–4):277 – 284, 2002. ISSN 0306-2619. doi: 10.1016/S0306-2619(02)00078-8.
- J. Hay and J. Davies. Calculation of the solar radiation incident on an inclined surface. In *Proc. 1st Canadian Solar Radiation Data Workshop*, volume 59, page , 1980.
- G. P. Henze, C. Felsmann, and G. Knabe. Evaluation of optimal control for active and passive building thermal storage. *International Journal of Thermal Sciences*, 43(2):173 – 183, 2004. ISSN 1290-0729. doi: 10.1016/j.ijthermalsci.2003.06.001.
- M. Jiménez, H. Madsen, and K. Andersen. Identification of the main thermal characteristics of building components using matlab. *Building and Environment*, 43(2):170–180, 2008. ISSN 03601323, 1873684x. doi: 10.1016/j.buildenv.2006.10.030.

- R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.
- H. Madsen. *Time Series Analysis*. Chapman & Hall, 2007.
- H. Madsen and J. Holst. Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings*, 22(1):67–79, 1995. ISSN 03787788.
- T. Mestekemper. *Energy demand forecasting and dynamic water temperature management*. PhD thesis, Bielefeld University, 2011.
- H. Nielsen, H. Madsen, and D. E. F. P. og Fordeling af El og Varme. Predicting the heat consumption in district heating systems using meteorological forecasts. Technical report, DTU IMM, 2000.
- H. A. Nielsen and H. Madsen. Modelling the heat consumption in district heating systems using a grey-box approach. *Energy & Buildings*, 38(1):63–71, 2006. ISSN 03787788. doi: 10.1016/j.enbuild.2005.05.002.
- T. Pedersen, P. Andersen, K. Nielsen, H. Starmose, and P. Pedersen. Using heat pump energy storages in the power grid. In *Control Applications (CCA), 2011 IEEE International Conference on*, pages 1106 –1111, sept. 2011. doi: 10.1109/CCA.2011.6044504.
- B. Perers, S. Furbo, J. Fan, E. Andersen, and Z. Chen. Solar combisystems with forecast control to increase the solar fraction and lower the auxiliary energy cost. In *ISES Solar World Congress 2011 Proceedings*, page , 2011. ISBN 978-3-9814659-0-7. Presented at: ISES Solar World Congress, SWC ; 30 : Kassel, Germany, 2011.
- T. Reddy, L. Norford, and W. Kempton. Shaving residential air-conditioner electricity peaks by intelligent use of the building thermal mass. *Energy*, 16 (7):1001 – 1010, 1991. ISSN 0360-5442. doi: 10.1016/0360-5442(91)90060-Y.
- D. Reindl, W. Beckman, and J. Duffie. Evaluation of hourly tilted surface radiation models. *Solar Energy*, 45(1):9 – 17, 1990. ISSN 0038-092X. doi: 10.1016/0038-092X(90)90061-G.
- J. Ruiz-Arias, H. Alsamamra, J. Tovar-Pescador, and D. Pozo-Vázquez. Proposal of a regressive model for the hourly diffuse solar radiation under all sky conditions. *Energy Conversion and Management*, 51(5):881–893, 2010. ISSN 01968904. doi: 10.1016/j.enconman.2009.11.024.
- P. d. Saint-Aubain. Adaptive load forecasting. Master’s thesis, Technical University of Denmark, 2011.
- Q. Zhou, S. Wang, X. Xu, and F. Xiao. A grey-box model of next-day building thermal load prediction for energy-efficient control. *International Journal of Energy Research*, 32(15):1418–1431, 2008. ISSN 1099-114X. doi: 10.1002/er.1458.

P A P E R G

Identifying suitable models for the heat dynamics of buildings

Authors:

Peder Bacher and Henrik Madsen

Published in:

Energy and Buildings, 2011, 43(7), pp. 1511-1522.

Revisions

Compared to the originally published paper the included paper has the following revisions:

- The estimates of the parameters related to the system and measurement noise have been included in Appendix G.9 together with a short discussion.

Abstract

The present paper suggests a procedure for identification of suitable models for the heat dynamics of a building. Such a procedure for model identification is essential for better usage of readings from smart meters, which is expected to be installed in almost all buildings in the coming years. The models can be used for different purposes, e.g. control of the indoor climate, forecasting of energy consumption, and for accurate description of energy performance of the building. Grey-box models based on prior physical knowledge and data-driven modelling are applied. This facilitates insight into otherwise hidden information about the physical properties of the building. A hierarchy of models of increasing complexity is formulated based on prior physical knowledge and a forward selection strategy is suggested enabling the modeller to iteratively select suitable models of increasing complexity. The performance of the models is compared using likelihood ratio tests, and they are validated using a combination of appropriate statistics and physical interpretation of the results. A case study is described in which a suitable model is sought after for a single storey 120 m² building. The result is a set of different models of increasing complexity, with which building characteristics, such as: thermal conductivity, heat capacity of different parts, and window area, are estimated.

G.1 Introduction

This paper describes a new method for obtaining detailed information about the heat dynamics of a building based frequent readings of the heat consumption, indoor temperature, ambient air temperature, and other climate variables. Such a method is considered to be of uttermost importance as a key procedure for better usage of readings from smart meters, which is expected to be installed in almost all buildings in the coming years. The method is based on a procedure for selecting a suitable model of the heat dynamics for a building. Rabl (1988) gives an overview of techniques for steady state and for dynamic analysis of energy use in a building, the latter implicate modelling of the heat dynamics of the building. Such dynamic models can be realized with a set of differential equations, as carried out by Sonderegger (1978) and Boyer et al. (1996). Parameter estimation in dynamical models is known as system identification and a survey of different approaches for buildings is found in Bloem (1994). The applied models in the present method are grey-box models, which consist of a set of continuous time stochastic differential equations coupled with a set of discrete time measurement equations. Grey-box modelling is well proven as a comprehensive and accurate method to model dynamical systems and thereby obtain knowledge of

the thermal properties of a building, see Madsen and Holst (1995), Andersen et al. (2000), and Jiménez et al. (2008). The problem of identifying a suitable model is both finding a model that is in agreement with the physical reality and finding a model, which has a complexity that is in agreement with the level of information embedded in data, which means that the model should neither be under-fitted nor over-fitted. The most suitable model is identified from a set of models of increasing complexity. A forward selection strategy is used, in which the modeller starts out with the simplest feasible model and iteratively selects models of increasing complexity. In each iteration the models are compared using likelihood-ratio tests and the models performances are evaluated. The selection procedure runs until no significant improvement of the model is found. See Pawitan (2001) and Madsen and Thyregod (2010) for an in-depth assessment of likelihood theory and model selection. The procedure is demonstrated by identifying a suitable model for a single storey 120 m² building. The building is part of the experimental distributed energy system, Syslab, at Risø DTU in Denmark. It is constructed of wood on the outside and plaster boards on the inside, with a layer of insulation wool in between. The data used spans 6 days and stems from a set of experiments for building energy performance, which was carried out in the winter period of 2009. It is thoroughly described in Bacher and Madsen (2010).

The remaining of the article is organized as follows. The applied grey-box modelling technique is described in Section G.2, and in Section G.3 the statistical test used for model selection is described. Then the suggested procedure for identifying a suitable model is outlined in Section G.4. The following section is devoted to a case-study, where the procedure is applied. It starts with a description of the building and the data, followed by an outline of the applied models and the selection, and ends with a discussion of the results. Finally a perspective of the applications are given in Section G.6 and the conclusions are drawn in Section G.7.

G.2 Grey-box models of a dynamic system

A grey-box model is established using a combination of prior physical knowledge and statistics, i.e. information embedded in data. The prior physical knowledge is formulated by a set of first-order stochastic differential equations, also called a stochastic linear state-space model in continuous time. The equations describe a lumped model of the heat dynamics of the building. It is emphasized that the physical interpretation of the parameters is dependent on how the building is divided into entities in the model.

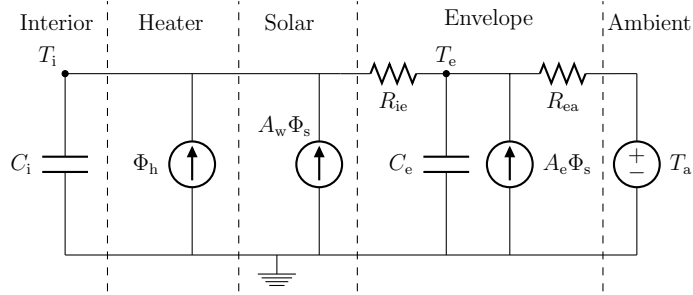


Figure G.1: RC-network of the model described by Equation (G.1) and (G.2). The model is divided into different parts indicating the corresponding part of the building.

An example of a feasible model is given here. It has two state variables, one describing the interior temperature T_i and one representing the temperature of the building envelope T_e . The first-order dynamics are represented by the stochastic differential equations

$$dT_i = \frac{1}{R_{ie}C_i}(T_e - T_i)dt + \frac{1}{C_i}\Phi_h dt + \frac{1}{C_i}A_w\Phi_s dt + \sigma_i d\omega_i \quad (\text{G.1})$$

$$dT_e = \frac{1}{R_{ie}C_e}(T_i - T_e)dt + \frac{1}{R_{ea}C_e}(T_a - T_e)dt + \sigma_e d\omega_e \quad (\text{G.2})$$

where t is the time, R_{ie} is the thermal resistance between the interior and the building envelope, R_{ea} is the thermal resistance between the building envelope and the ambient air, C_i is the heat capacity of the interior, C_e is the heat capacity of the building envelope, Φ_h is the energy flux from the heating system, A_w is the effective window area, Φ_s is the energy flux from solar radiation, T_a is the ambient air temperature, $\{\omega_{i,t}\}$ and $\{\omega_{e,t}\}$ are standard Wiener processes, and σ_i^2 and σ_e^2 are the incremental variances of the Wiener processes. The model can be represented with the RC-network depicted in Figure G.1, where the model is divided into different parts to show the corresponding parts of the building.

The physical model part is coupled with the data-driven model part with which the information embedded in observed data is used for parameter estimation. The data-driven part in the considered example is represented by the discrete time measurement equation

$$Y_k = T_{ik} + e_k \quad (\text{G.3})$$

where k is the point in time t_k of a measurement, Y_k is the measured interior temperature and e_k is the measurement error, which is assumed to be a Gaussian white noise process with variance σ^2 . This assumption enables evaluation and tests of the performance of the model, since if it is met, this indicates that the physical model is consistent with the observed heat dynamics of the building.

G.2.1 Maximum likelihood estimation of parameters

Given a grey-box model, as described above, maximum likelihood estimates of the parameters can be obtained. Let the observations be represented by

$$\mathcal{Y}_N = [Y_N, Y_{N-1}, \dots, Y_1, Y_0] \quad (\text{G.4})$$

then the likelihood function is the joint probability density

$$L(\theta; \mathcal{Y}_N) = \left(\prod_{k=1}^N p(Y_k | \mathcal{Y}_{k-1}, \theta) \right) p(Y_0 | \theta) \quad (\text{G.5})$$

where $p(Y_k | \mathcal{Y}_{k-1}, \theta)$ is a conditional density denoting the probability of observing Y_k given the previous observations and the parameters θ , and where $p(Y_0 | \theta)$ is a parameterization of the starting conditions. The maximum likelihood estimates of the parameters are then found by

$$\hat{\theta} = \arg \max_{\theta} \{L(\theta; \mathcal{Y}_N)\} \quad (\text{G.6})$$

Due to the previously mentioned assumptions about the noise process and the fact that the model is linear, it follows that the conditional densities in Equation (G.6) are Gaussian densities. Since the conditional densities are Gaussian a Kalman filter can be used to calculate the likelihood function, and an optimization algorithm can be applied to maximize it, thereby calculating the maximum likelihood estimates, see Kristensen et al. (2004) for a detailed discussion. This is implemented in the computer software CTSM, which has been used for carrying out the parameter estimation, see more about the software at ¹ and in Kristensen and Madsen (2003).

G.3 A statistical test for model selection

Statistical tests can be utilized in the search for the most suitable model. If a model is a sub-model of larger model, then a likelihood ratio test can determine if the larger model performs significantly better than the sub-model. Using a sequence of such tests a strategy for selection of the best model can be evolved.

G.3.1 Likelihood ratio tests

Let a model have parameters $\theta \in \Omega_0$ where $\Omega_0 \in \mathbb{R}^r$ is the parameter space and $r = \dim(\Omega_0)$ is the number of parameters in the model. Let a larger model have

¹www.imm.dtu.dk/~ctsm

parameters $\theta \in \Omega$ where $\Omega \in \mathbb{R}^m$ and $\dim(\Omega) = m$, and assume that

$$\Omega_0 \subset \Omega \quad (\text{G.7})$$

i.e. the first model is a sub-model of the second model and $r < m$.

The likelihood ratio test

$$\lambda(\mathcal{Y}_N) = \frac{\sup_{\theta \in \Omega_0} L(\theta; \mathcal{Y}_N)}{\sup_{\theta \in \Omega} L(\theta; \mathcal{Y}_N)} \quad (\text{G.8})$$

where \mathcal{Y}_N is the observed values, can then be used to test the hypothesis

$$H_0 : \theta \in \Omega_0 \quad \text{vs.} \quad H_a : \theta \in \Omega \setminus \Omega_0 \quad (\text{G.9})$$

since under H_0 the test statistic $-2\log(\lambda(\mathcal{Y}_N))$ converges to a χ^2 distributed random variable with $(m - r)$ degrees of freedom as the number of samples in \mathcal{Y}_N goes to infinity. If H_0 is rejected then the likelihood of the larger model is significantly higher than the likelihood of the sub-model, and it is concluded that \mathcal{Y}_N is more likely to be observed with the larger model. Hence the larger model is needed over the sub-model to describe the information embedded in data. For more details see Madsen and Thyregod (2010).

G.3.2 Forward selection

In a forward selection procedure the modeller starts with the smallest feasible model and then in each step extends the model with the part that gives the lowest p -value, i.e. the most significant improvement. The possible candidates for improvement that are selected in each iteration are the smallest possible extensions to the current model. The procedure stops when no extensions to the model yields a p -value below a pre-specified limit, usually set to 5%.

G.4 Model selection procedure

Different strategies for identifying a suitable model is proposed in the literature and finding an appropriate strategy depends on the specific modelling setup. An purely algorithmic and exhaustive selection procedure is seldomly feasible, hence iterative methods, in which the modeller is partly involved in the selection, are commonly applied. Here, a forward selection procedure is suggested for identification of a suitable model for the heat dynamics. It is based on likelihood ratio testing, which is described in Section G.3.1.

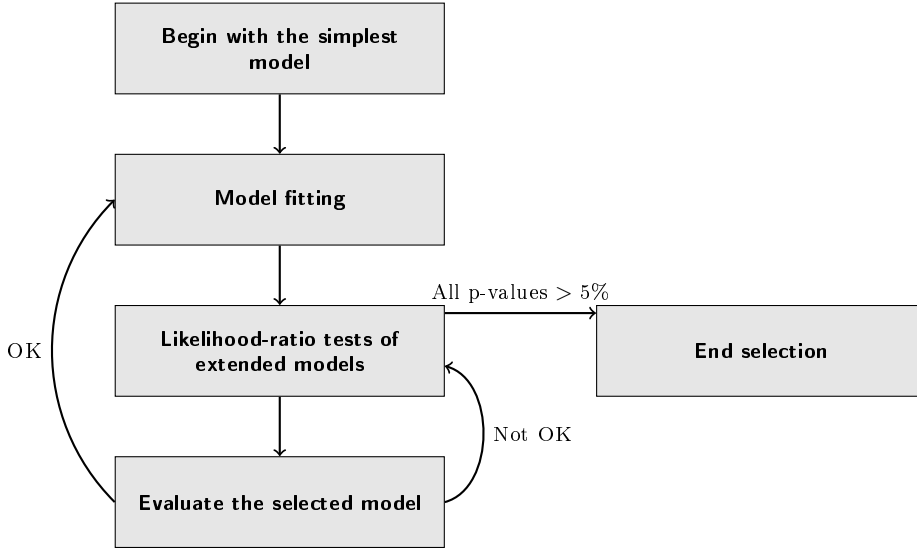


Figure G.2: Illustration of the model selection procedure

G.4.1 Model selection

The procedure starts by a formulation of the simplest feasible model having parameter space Ω_m and a full model with parameter space Ω_{full} , such that

$$\Omega_m \subset \Omega_{full}. \quad (G.10)$$

Within this range a set of models can be formed, in which a suitable model is to be identified. A suitable model is a sufficient model, which is the smallest model that describes all information embedded in the data (Madsen and Thyregod, 2010). The selection is initiated with the simplest model and then extensions of the model are iteratively added. The selection stops when all of the extensions to the selected model, gives a likelihood-ratio test p -value above the pre-specified limit. Hence the procedure will stop with a model from which no larger model can be found, with which it is significantly more likely to observe the data. As mentioned above a purely algorithmic procedure is not possible, hence the modeller must be involved to evaluate the models estimated in each iteration. The evaluation is carried out by analyzing the properties of residuals and parameter estimates, and if some of the properties are not in line with the assumptions and physical reality, then the modeller may have to influence the choice of model. The procedure is illustrated in Figure G.2 and the steps are

Model fitting The models which are extended from the current model are

fitted to the data by maximum likelihood estimation of the parameters.

Likelihood-ratio tests Calculate the likelihood-ratio test statistic for the current model versus each of the extended models. Stop if none of the tests have a p -value below 5% and use the current model as the final model, else then select the extended model which yield the lowest p -value.

Evaluate The modeller evaluates the selected extended model. If the result is satisfactory the model is kept and next iteration can be started; if not, the previous step should be repeated to select another extension.

If two extensions show an almost identical improvement, i.e. the p -values of the tests are nearly equal, the selection can be branched and extensions with different parts examined separately. The procedure will then end with several models, which cannot be tested directly against each other, and it is then up to the modeller to decide which should be preferred. This should be done by comparing the likelihoods, where if two models have almost equal likelihoods the smaller model should be preferred, and furthermore by an evaluation of the residuals and parameter estimates. It can also happen that several models have only marginal difference in performance and that each of them can be considered to be a sufficient model.

G.4.2 Model evaluation

In each step the selected model must be evaluated. This serves both to check if the model meet the assumptions and if it gives reasonable estimates from a physical point of view. Furthermore the evaluation can reveal model deficiencies from which it can be learned which parts of the model should be further elaborated. The evaluation should consist of the following:

- The assumption of white noise residuals should be inferred upon using the auto-correlation function (ACF) and the cumulated periodogram (CP), which can also reveal how well dynamics on different timescales are modelled.
- Plots of the inputs, outputs, and residuals. These plots can be used to understand which effects are not described properly by the model.
- Evaluation of the estimated physical parameters. Clearly the results should be consistent among different models, e.g. estimate of the thermal resistance of the building envelope should not change significantly among the models. Furthermore the modeller have to judge if the results are consistent with reality.



Figure G.3: (a) the north facade and (b) is the south facade of the building.

G.5 Case study: model identification for a building

The method is demonstrated by applying it to identify a suitable model for a building. The building, named FlexHouse, is part of the experimental energy system Syslab, at Risø DTU in Denmark. It is well suited for such experiments since it has a controllable electrical heating system. Measurements consisting of five minute values over a period of six days are used, for further details see Bacher and Madsen (2010), in which a thorough description of the experiments and data is given. This section starts with a description of the building and measurement equipment, then the data is presented, followed by an outline of the considered models, and finally the model identification and evaluation is given.

G.5.1 Description of the building and measurement equipment

The outer walls of the building are constructed of wood on the outside and plaster boards on the inside, with a layer of insulation wool in between. An image of the north facade and an image of the south facade of the building can be seen in Figure G.3. The building rests on piles, leaving an air gap between the ground and the building. The roof is flat and covered with roofing felt. The dimensions of the floor plan is approximately 7.5 times 16 meters. In Figure G.4 the floor plan of the building is shown. A server system is installed in the building, which can control the electrical heaters located as indicated on the floor

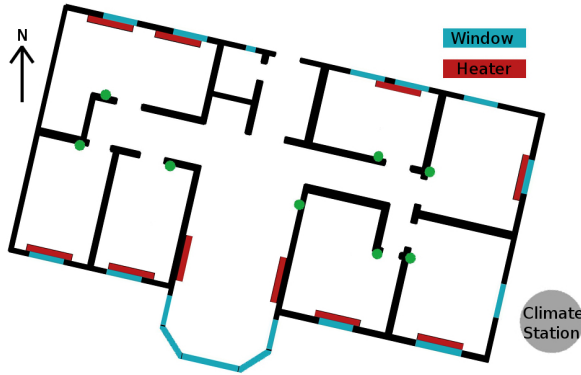


Figure G.4: The floor plan of the building.

plan. To measure the indoor temperature Hobo U12-012 Temp/RH/Light/Ext sensors mounted on a small piece of wood was hanged freely in the middle of each room. A small climate station is located two meters east of the building, the position relative to the building is indicated in Figure G.4.

G.5.2 Data

The present study is based on data collected during a series of experiments carried out in February to April 2009. The following time series consisting of five minute average values are used:

y ($^{\circ}\text{C}$) A single signal representing the indoor temperature. It is formed as the first principal component of the measurements of the indoor temperature from the Hobo sensors.

T_a ($^{\circ}\text{C}$) Observed ambient air temperature at the climate station.

Φ_h (kW) Total heat input from the electrical heaters in the building.

Φ_s (kW/m^2) The global irradiance measured at the climate station.

Plots of the time series can be found in Figure G.5. The controlled heat input is a pseudo-random binary sequence (PRBS), which has white noise properties and no correlation with the other inputs. It is designed to excite the heat dynamics at several ranges of frequencies in which the time constants of the building is expected to be, such that the information embedded in data is optimized

for estimation of the heat dynamic properties of the building, see Madsen and Schultz (1993).

G.5.3 Applied models

The proposed procedure is such, that the modeller starts with the simplest model and iteratively selects more complex models. This implies fitting a set of models from the simplest model to the most feasible complex model, denoted the full model. In this section the set of applied models and the result of the iterative selection procedure is described. All the models are grey-box models, in which the physical part is stochastic linear state-space model (as presented in Section G.2) and where the dynamics of the states can be written

$$d\mathbf{T} = \mathbf{A}\mathbf{T}dt + \mathbf{B}\mathbf{U}dt + d\boldsymbol{\omega} \quad (\text{G.11})$$

where \mathbf{T} is the state vector and \mathbf{U} is the input vector, and none of the state variables or input variables are in \mathbf{A} or \mathbf{B} which only consist of parameters. All the considered models have an input vector with three inputs

$$\mathbf{U} = [T_a, \Phi_s, \Phi_h]^T \quad (\text{G.12})$$

All the models are lumped, but with a different structure, which implies that a given parameter does not necessarily represent the same physical entity in each model. For example the parameter C_i is representing the heat capacity of the entire building in the simplest model, whereas this heat capacity is divided into five heat capacities in the full model, in which the parameter C_i represents the heat capacity of the indoor air. This is elaborated further in Section G.5.5.2, where the parameter estimates for the models are presented. Furthermore it should be kept in mind that these models are linear approximations to the real system.

In the following sections the full and the simplest model are described, since they represent the range of applied models. First the full model is outlined to give a complete overview of all the individual parts, which is included in the models. Then the simplest model is presented, since it is the first model applied in the selection procedure and furthermore it illustrates how the models are lumped. Each model is named from its state vector and where needed a few parameter names. See G.8 for a list of RC-networks corresponding to all applied models.

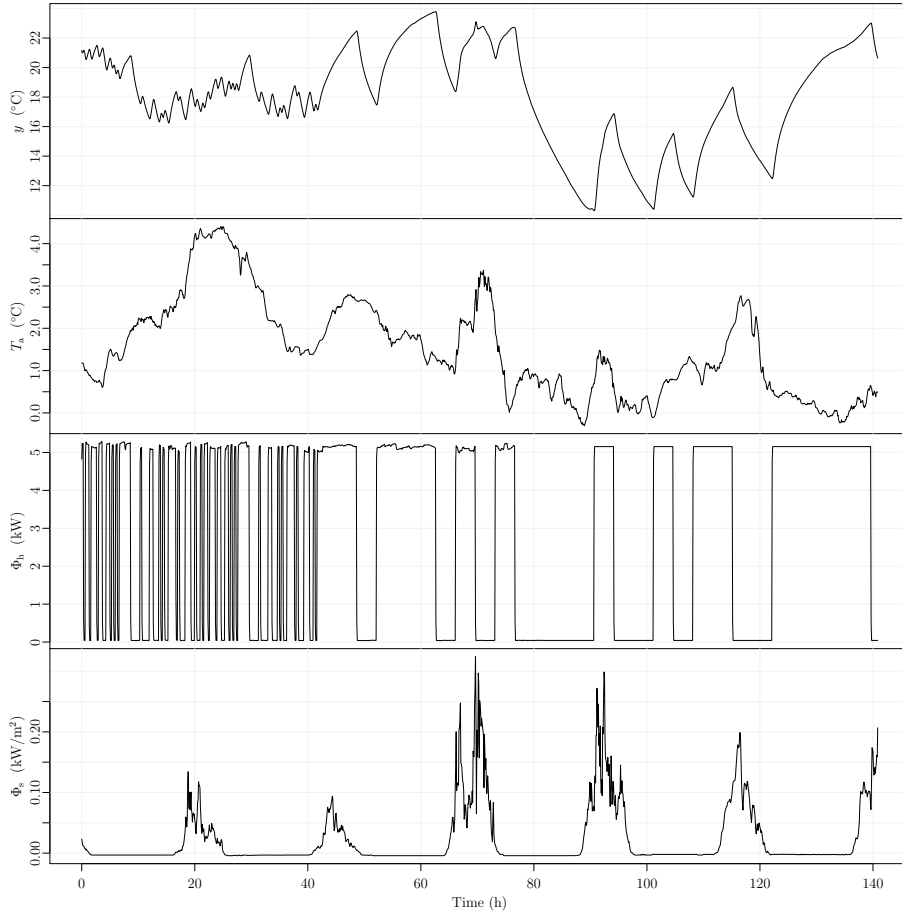


Figure G.5: The data set. From the top, the first plot shows the observed indoor temperature y , the second shows the ambient air temperature T_a , followed by a plot of the heat input Φ_h , and finally the lower plot shows the global irradiance Φ_s .

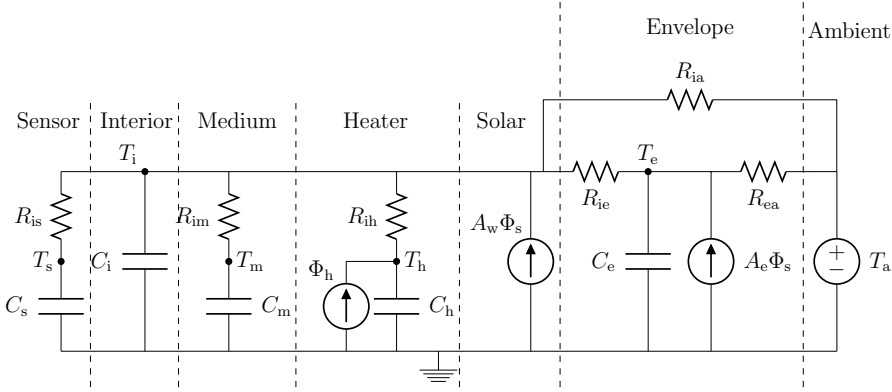


Figure G.6: The full model $TiTmTeThTsAeRia$ with the individual model parts indicated. This model includes all parts which is included in any of the applied models.

G.5.3.1 The full model $TiTmTeThTsAeRia$

The RC-network of the full model, which is the most complex model applied, is illustrated in Figure G.6. This model includes all the individual parts of the building, which it is found feasible to include in linear models, with the current available data. The individual model parts are indicated on the figure. The model parts are:

Sensor The temperature sensors are modelled with a heat capacity and a thermal resistance to the interior.

Interior In the full model the interior is considered to be the indoor air (again remember that, since the models are lumped models, the building part represented by “Interior” is mostly different for each model) and it is modelled as a heat capacity connected to other parts by thermal resistances.

Medium A thermal medium inside the building is the interior walls and furniture, which is modelled with a heat capacity and a thermal resistance to the interior.

Heater The heaters are modelled by a heat capacity and a thermal resistance to the interior.

Solar The heat input from solar radiation is modelled by the global irradiance multiplied with the effective window area.

Envelope The building envelope is modelled with a heat capacity and thermal resistances to both the interior and the ambient. A thermal resistance directly coupled to the ambient is also included.

Ambient The ambient is represented by the observed ambient air temperature.

The full model includes five state variables, that each represents the temperature in a part of the building, and they are:

T_s The temperature of the sensor, which for the full model is used as the model output, i.e. Y_k in the measurement equation (Equation (G.19)).

T_i The temperature of the interior, i.e. the indoor air.

T_m The temperature of an interior thermal medium, i.e. interior walls and furniture.

T_h The temperature of the heaters.

T_e The temperature of the building envelope.

The parameters of the model represent different thermal properties of the building. This includes thermal resistances:

R_{is} between the interior and the sensor,

R_{im} between the interior and the interior thermal medium,

R_{ih} between the heaters and the interior,

R_{ia} between the interior and the ambient,

R_{ie} between from the interior and the building envelope,

R_{ea} between the building envelope and the ambient.

The heat capacities of different parts of the building are represented by:

C_s for the temperature sensor,

C_i for the interior,

C_m for the interior walls and furniture,

C_h for the electrical heaters,

C_e for the building envelope.

Finally two coefficients are included, each representing an estimate of an effective area in which the energy from solar radiation enters the building. They are:

A_w The effective window area of the building.

A_e The effective area in which the solar radiation enters the building envelope.

The set of stochastic differential equations describing the heat flows in the full model are

$$dT_s = \frac{1}{R_{is}C_s}(T_i - T_s)dt + \sigma_s d\omega_s \quad (G.13)$$

$$dT_i = \frac{1}{R_{is}C_i}(T_s - T_i)dt + \frac{1}{R_{im}C_i}(T_m - T_i)dt + \frac{1}{R_{ih}C_i}(T_h - T_i) \quad (G.14)$$

$$\frac{1}{R_{ie}C_i}(T_e - T_i)dt + \frac{1}{R_{ia}C_i}(T_a - T_i)dt + \frac{1}{C_i}A_w\Phi_s dt + \sigma_i d\omega_i \quad (G.15)$$

$$dT_m = \frac{1}{R_{im}C_m}(T_i - T_m)dt + \sigma_m d\omega_m \quad (G.16)$$

$$dT_h = \frac{1}{R_{ih}C_h}(T_i - T_h)dt + \frac{1}{C_h}\Phi_h dt + \sigma_h d\omega_h \quad (G.17)$$

$$dT_e = \frac{1}{R_{ie}C_e}(T_i - T_e)dt + \frac{1}{R_{ea}C_e}(T_a - T_e)dt + \frac{1}{C_e}A_e\Phi_s dt + \sigma_e d\omega_e \quad (G.18)$$

and the measurement equation is

$$Y_k = T_{s,k} + e_k \quad (G.19)$$

since the observed temperature is encumbered with some measurement error.

G.5.3.2 The simplest model T_i

The simplest model considered is illustrated by the RC-network in Figure G.7. The model has one state variable T_i and the following parameters:

R_{ia} The thermal resistance from the interior to the ambient.

C_i The heat capacity of the entire building, including the indoor air, interior walls, furniture etc., and the building envelope.

A_w is the effective window area of the building.

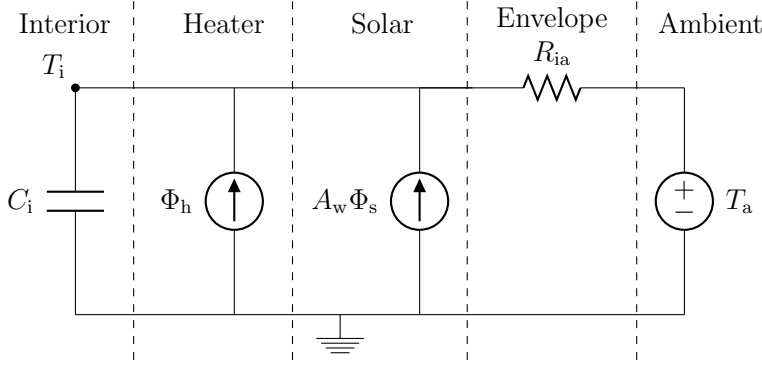


Figure G.7: RC-network of the model T_i , which is the simplest feasible model.

The stochastic differential equation describing the heat flow is

$$\frac{dT_i}{dt} = \frac{1}{R_{ia}C_i}(T_a - T_i) + \frac{1}{C_i}A_w\Phi_s + \frac{1}{C_i}\Phi_h + \sigma_i \frac{d\omega_i}{dt} \quad (\text{G.20})$$

and the measurement equation is

$$Y_k = T_{i,k} + e_k \quad (\text{G.21})$$

Note the differences in representation of the building parts between the simplest and full model, e.g. R_{ia} represent the thermal resistance of the building envelope in the simplest model, whereas this is represented by a coupling of R_{ia} , R_{ie} , and R_{ea} in the full model.

G.5.4 Model identification

The identification procedure is applied to find a sufficient model in the set of models ranging from T_i to $T_iTmTeThTsAeRia$. The log-likelihood of each model, which is fitted, is listed in Table G.1 ordered by the iterations of the model selection. The procedure begins with the simplest model. Then in the first iteration four extended models are fitted and T_iTh is selected since it has the highest log-likelihood, hence the lowest p -value of the likelihood-ratio tests (the four models have the same number of parameters). The selection procedure is carried out until no significant extension can be found, which occurs in iteration number five. During each iteration the current selected model is evaluated, see Section G.5.5. It is found that the models selected in each iteration are all satisfying the evaluation with respect to improvement of the results etc. In Table G.2 the result of likelihood-ratio tests for model expansion in each iteration is

Iteration	Models			
Start	<i>Ti</i>			
$l(\theta; \mathcal{Y}_N)$	2482.6			
m	6			
1	<i>TiTe</i>	<i>TiTm</i>	<i>TiTs</i>	<i>TiTh</i>
	3628.0	3639.4	3884.4	3911.1
	10	10	10	10
2	<i>TiThTs</i>	<i>TiTmTh</i>	<i>TiTeTh</i>	
	4017.0	5513.1	5517.1	
	14	14	14	
3	<i>TiTeThRia</i>	<i>TiTeThAe</i>	<i>TiTmTeTh</i>	<i>TiTeThTs</i>
	5517.3	5520.5	5534.5	5612.4
	15	15	18	18
4	<i>TiTeThTsRia</i>	<i>TiTmTeThTs</i>	<i>TiTeThTsAe</i>	
	5612.5	5612.9	5614.6	
	19	22	19	
5	<i>TiTmTeThTsAe</i>	<i>TiTeThTsAeRia</i>		
	5614.6	5614.7		
	23	20		

Table G.1: Log-likelihood $l(\theta; \mathcal{Y}_N)$ for the fitted models ordered by iterations of the model selection procedure and in each row by log-likelihood. In each iteration the extended model with highest log-likelihood is selected, which is the rightmost models in the table. The number of estimated parameters for each model is indicated by m .

listed. Clearly, the expansions carried out in the first three iterations indicate very significant improvements of the model. In iteration four, the improvement is still below 5%, whereas no significant improvement is found in iteration five. The procedure thus ends with *TiTeThTsAe* as a sufficient model, which is illustrated by the RC-network in Figure G.8.

G.5.5 Model evaluation

In the following the selected models are evaluated as outlined in Section G.4.2.

Iteration	Sub-model	Model	$m - r$	$-2\log(\lambda(y))$	p -value
1	T_i	T_iTh	4	2857	$< 10^{-16}$
2	T_iTh	T_iTeTh	4	3212	$< 10^{-16}$
3	T_iTeTh	$T_iTeThTs$	4	190.5	$< 10^{-16}$
4	$T_iTeThTs$	$T_iTeThTsAe$	1	4.5	0.035
5	$T_iTeThTsAe$	$T_iTeThTsAeRia$	1	0.12	0.73

Table G.2: Tests carried out in the model selection procedure.

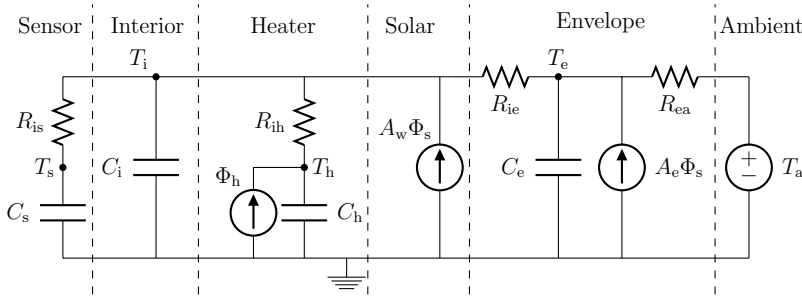


Figure G.8: The final selected model $T_iTeThTsAe$ with the individual model parts indicated.

G.5.5.1 Residuals

Plots of output, inputs, and residuals for each model can be seen in Figure G.9. For each model the auto-correlation function (ACF) of the residuals is plotted in Figure G.10 and the cumulated periodogram (CP) in Figure G.11. It is seen directly from the plot of the residuals from the simplest model T_i , that they do not have white noise properties and that they are not independent of the inputs. The ACF of the residuals also clearly show a high lag dependency, and the CP reveals that the model is not detailed enough to describe the dynamics. Examining the plot of the residuals for the model selected in the first iteration, T_iTh , it is seen that the level of the residuals is reduced compared to the residuals for T_i . The ACF and CP indicate that the assumption of white noise residuals is not fulfilled. From the plot of the residuals for the model selected in the second iteration, T_iTeTh , it is seen that the level of the residuals is reduced dramatically, but that some dependency of the inputs is still seen, mostly from the solar irradiance. The ACF reveals that the characteristics of the residuals are much closer to white noise, which is also seen from the CP, indicating that the model now describes the heat dynamics of the building quite well. The plot of the residuals, ACF, and CP for the model selected in the third iteration

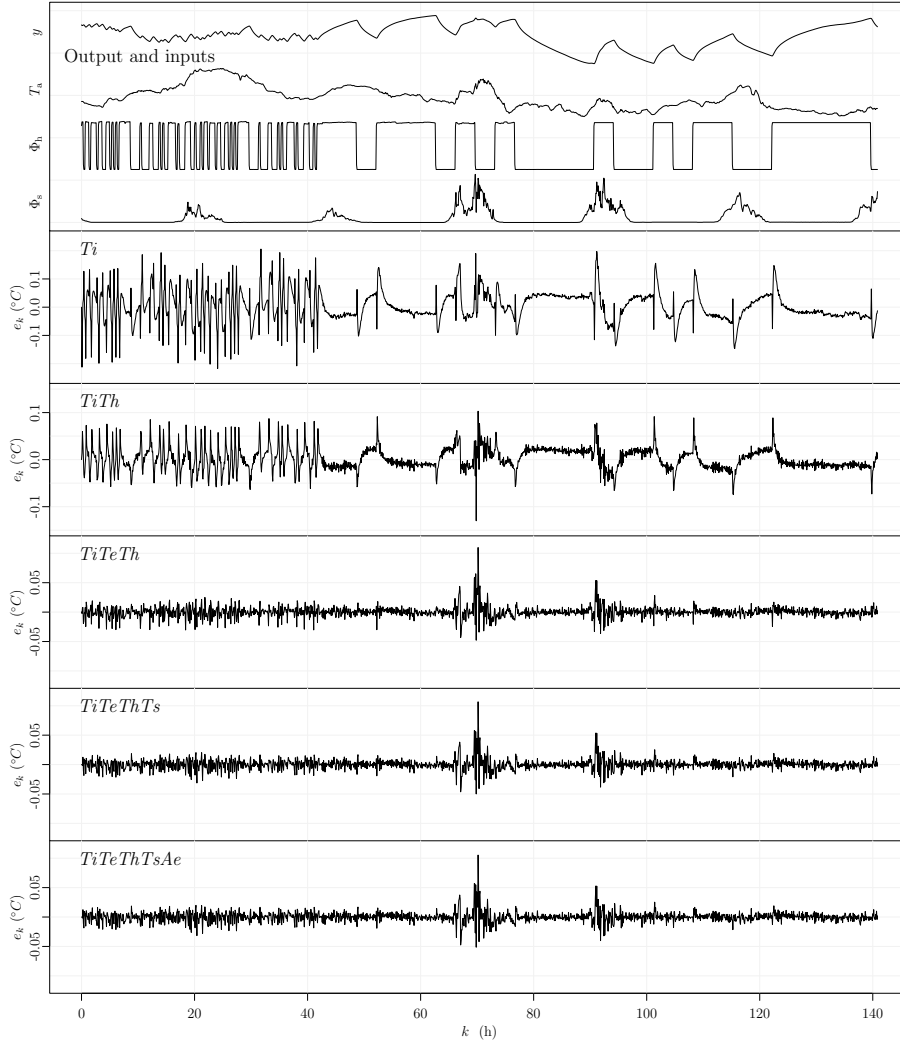


Figure G.9: The upper plot is of the output and inputs, and the following plots are of the residuals for each of the selected models. On each plot of the residuals the model name is indicated.

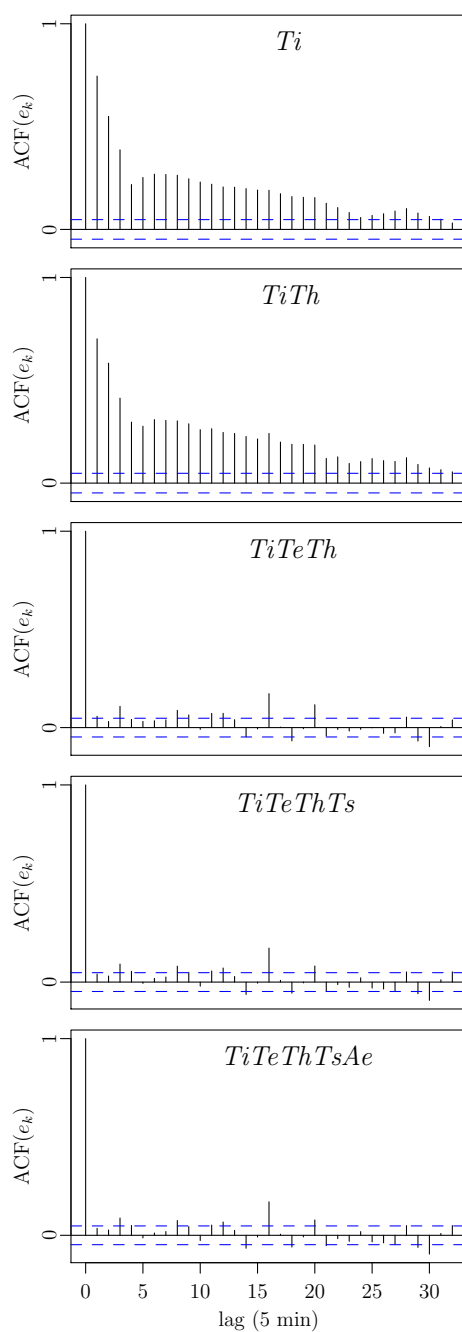


Figure G.10: The auto-correlation function of the residuals for each of the selected models.

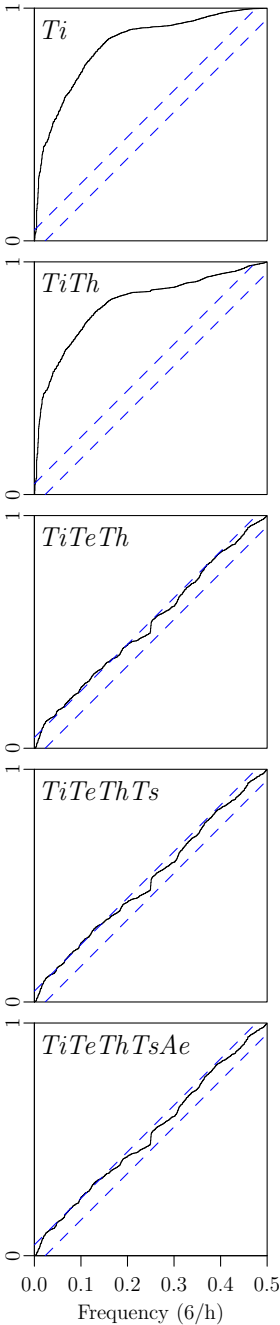


Figure G.11: The cumulated periodogram of the residuals for each of the selected models.

$TiTeThTs$, reveals only slight improvements compared to the previous model. Finally the plots for the final selected model $TiTeThTsAe$, show that almost no differences can be observed from the previous model. The highest level of error can be observed where the solar irradiance is high, hence it is found that further improvement of the model should be focused on the part in which the solar radiation enters the building, or alternatively in letting the incremental variance of the Wiener process depend on the solar radiation.

G.5.5.2 Parameter estimates

The parameter estimates of the selected models are evaluated in the following. The estimates are presented in Table G.3 together with the time constants calculated for each of the selected models. The total heat capacity and thermal resistance of the building envelope estimated by the selected models are presented in Table G.4. As found by evaluating the residuals, see previous section, the models Ti and $TiTh$ do not describe the dynamics of the system very well, which implies that the estimates of the heat capacities are not reliable. Esti-

Model	Ti	$TiTh$	$TiTeTh$	$TiTeThTs$	$TiTeThTsWithAe$
C_i	2.07	1.36	1.07	0.143	0.0928
C_e	-	-	2.92	3.24	3.32
C_h	-	0.309	0.00139	0.321	0.889
C_s	-	-	-	0.619	0.0549
R_{ia}	5.29	5.31	-	-	-
R_{ie}	-	-	0.863	0.909	0.897
R_{ea}	-	-	4.54	4.47	4.38
R_{ih}	-	0.639	93.4	0.383	0.146
R_{is}	-	-	-	0.115	1.89
A_w	7.89	6.22	5.64	6.03	5.75
A_e	-	-	-	-	3.87
τ_1	10.9	0.16	0.129	0.0102	0.0102
τ_2	-	8.9	0.668	0.105	0.105
τ_3	-	-	18.4	0.786	0.788
τ_4	-	-	-	19.6	19.3

Table G.3: The estimated parameters. The heat capacities, C_x , are in [kWh/°C]. The thermal resistances, R_{xx} , are in [°C/kW]. The areas, A_x , are in [m²]. The time constants, τ_x , are in hours. Note that the physical interpretation for many of the parameters is different for each model.

Model	T_i	T_iTh	T_iTeTh	$T_iTeThTs$	$T_iTeThTsAe$
C_{total}	2.07	1.67	3.99	4.32	4.36
$R_{envelope}$	5.29	5.31	5.40	5.38	5.28
α_{UA}	1.55	1.55	1.52	1.53	1.55

Table G.4: The total heat capacity [kWh/°C] and thermal resistance [°C/kW] of the building envelope estimated by the selected models. The UA-values α_{UA} are in [W/(°Cm²)].

mates of the heat capacities found by the three larger models are more credible, especially it is seen that the time constants are almost equal, indicating that the models comprise the same dynamics. The exact physical interpretation of the smaller heat capacities C_i , C_h , and C_s cannot be given, but it is noted that their sum, for each of the three larger models, is quite close ranging from 1.03 to 1.08 [kWh/°C].

The estimated total thermal resistance of the building envelope and thereby the UA-values is quite similar for all models, as seen in Table G.4.

G.6 Applications

Identification of a suitable model of the heat dynamics of a building based on frequent readings of heat consumption, indoor temperature, ambient air temperature, and climate variables, will be very useful for different purposes. Important fields of application are:

Accurate description of energy performance of the building An energy signature of buildings can provide important information for energy- and cost effective improvements of the building. The most effective actions to be taken for an individual building can be identified. Furthermore the heat consumption due to physical effects, such as a poor isolated building envelope, can be separated from behavioral effects, e.g. a high indoor temperature.

Forecasting of energy consumption for heating Forecasting of energy consumption for heating can be used for integration of large amounts of renewable energy, such as wind- and solar energy. Implementation of electrical heating with hot water tanks for heat storage in individual houses can be profitable in the near future. Knowledge of the heat dynamics of buildings is essential to forecasting and control of such systems.

Indoor climate control Control of the indoor temperature, ventilation etc. to provide a good indoor climate conditions can be carried out with methods which include models of the heat dynamics. The models can also be extended to include the effect of wind and thereby provide information of the air tightness of buildings.

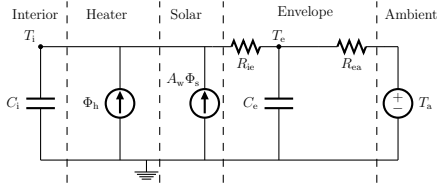
G.7 Conclusion

A procedure for identification of the most suitable models for the heat dynamics of a building has been described and applied on the basis of data from an experiment carried out in February 2009. The procedure is based on likelihood-ratio testing combined with a forward selection strategy. The proposed models are grey-box models, where a combination of prior physical knowledge and data-driven modelling is utilized. The data used for the modelling consist of: climate data measured at the location, measurements of the indoor temperature, and a PRBS controlled heat input.

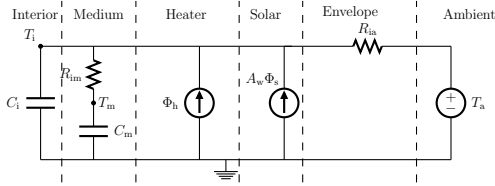
The results of the identification procedure are evaluated and discussed, both in a statistical and physical context. The evaluation reveal that the selected model meet the assumptions of white noise residuals, hence it can be applied to give reliable estimates consistent with reality and the results are statistically validated. Furthermore model deficiencies are pointed out, from which further advancement of the model should be pursued. For the considered building this is primarily on the model part where the solar radiation input enters the building.

It has been shown that the method is able to provide rather detailed knowledge of the heat dynamics of the building. This includes for instance the thermal resistance of the envelope and thereby the UA-value, parameters describing the capabilities for storing heat, and the time constants of the building.

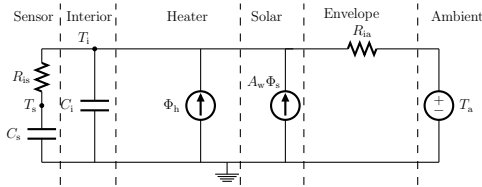
G.8 RC-networks of applied models



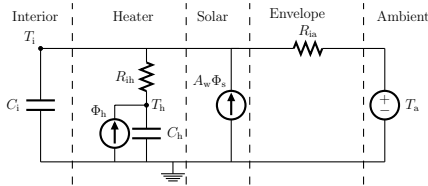
(a) RC-network network of $T_i T_e$.



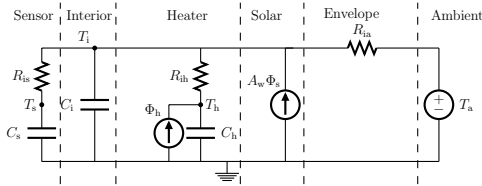
(b) RC-network network of $T_i T_m$.



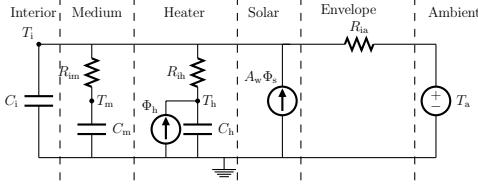
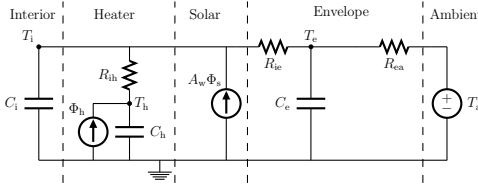
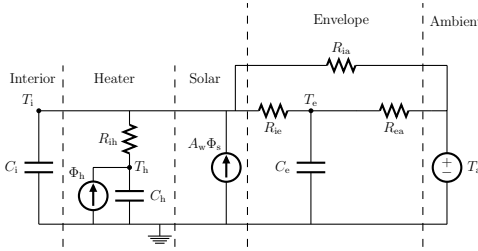
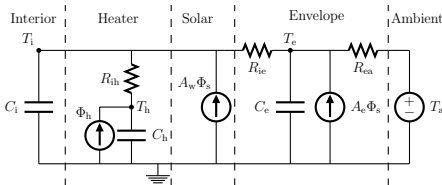
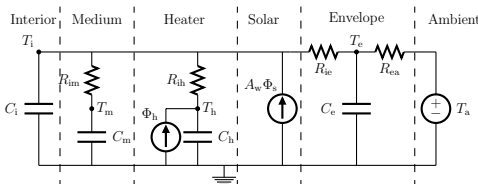
(c) RC-network network of $T_i T_s$.

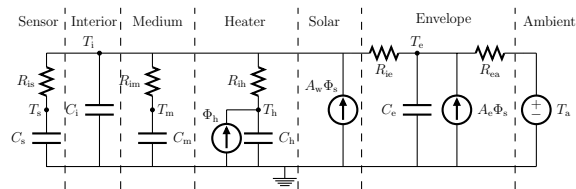
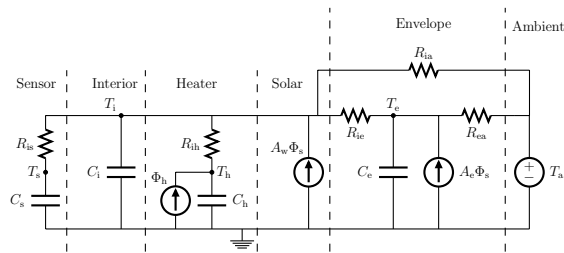
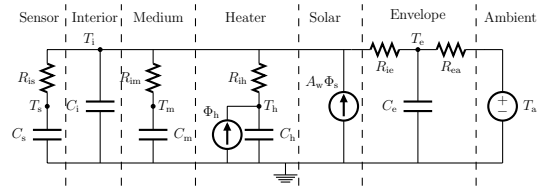
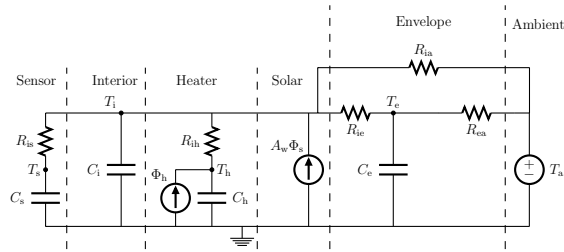
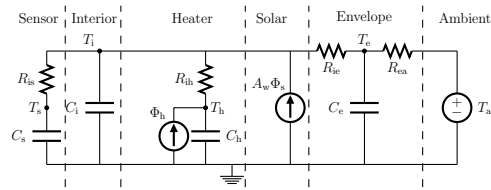


(d) RC-network network of $T_i T_h$.



(e) RC-network network of $T_i T_h T_s$.

(a) RC-network network of $T_i T_m T_h$.(b) RC-network network of $T_i T_e T_h$.(c) RC-network network of $T_i T_e T_h R_{ia}$.(d) RC-network network of $T_i T_e T_h A_e$.(e) RC-network network of $T_i T_m T_e T_h$.



G.9 Estimates of system and observation noise parameters

The estimates of the scaling of the increments in the Wiener process in the system equations together with the standard deviation of the observation noise are listed in Table G.5. Considering the level of the estimated system noise for *Ti* and *TiTh* compared to the level of the estimated system noise for each of the three larger models, it should intuitively be higher, since *Ti* is a submodel of the larger models. However, since the residuals for *Ti* and *TiTe* are not white noise, then the system noise is almost surely not a Wiener process. Hence the definition of the parameters are not valid and a direct comparison not possible. For the three larger models *TiTeTh*, *TiTeThTs*, and *TiTeThTsWithAe* the estimated level of system noise is high compared to the level of measurement noise. This indicates that not all information is described by the model, hence that a more complicated and possibly non-linear model could be identified as more suitable than *TiTeThTsWithAe*.

Model	<i>Ti</i>	<i>TiTh</i>	<i>TiTeTh</i>	<i>TiTeThTs</i>	<i>TiTeThTsWithAe</i>
σ_i	0.19	$4.9 \cdot 10^{-6}$	0.017	0.13	0.18
σ_e			0.27	0.28	0.28
σ_h		1.4	$4.0 \cdot 10^{-5}$	$7.2 \cdot 10^{-5}$	$2.7 \cdot 10^{-5}$
σ_s				$3.8 \cdot 10^{-13}$	$2.0 \cdot 10^{-12}$
σ	$1.7 \cdot 10^{-11}$	$1.7 \cdot 10^{-11}$	$2.4 \cdot 10^{-6}$	$1.1 \cdot 10^{-6}$	$1.4 \cdot 10^{-6}$

Table G.5: The estimated values of system noise and measurement noise related parameters. The unit of the system noise parameters $\sigma_i, \sigma_e, \sigma_h, \sigma_s$ is $^{\circ}\text{C}/\sqrt{\text{h}}$ and the unit of the measurement noise σ is $^{\circ}\text{C}$.

References

- K. K. Andersen, H. Madsen, and L. H. Hansen. Modelling the heat dynamics of a building using stochastic differential equations. *Energy and Buildings*, 31 (1):13–24, 2000. ISSN 03787788.
- P. Bacher and H. Madsen. Experiments and data for building energy performance analysis : Financed by the danish electricity saving trust. Technical report, DTU Informatics, Building 321, Kgs. Lyngby, 2010.

- J. J. Bloem, editor. *System Identification Applied to Building Performance Data*. CEC-EUR 15885 EN, 1994.
- H. Boyer, J. Chabriat, B. GrondinPerez, C. Tourrand, and J. Brau. Thermal building simulation and computer generation of nodal models. *BUILDING AND ENVIRONMENT*, 31(3):207–214, MAY 1996. ISSN 0360-1323.
- M. Jiménez, H. Madsen, J. Bloem, and B. Dammann. Estimation of non-linear continuous time models for the heat exchange dynamics of building integrated photovoltaic modules. *Energy & Buildings*, 40(2):157–167, 2008. ISSN 03787788.
- N. R. Kristensen and H. Madsen. Continuous time stochastic modelling, CTSM 2.3 - mathematics guide. Technical report, DTU, 2003.
- N. R. Kristensen, H. Madsen, and S. B. Jørgensen. Parameter estimation in stochastic grey-box models. *Automatica*, 40(2):225 – 237, 2004. ISSN 0005-1098. doi: DOI:10.1016/j.automatica.2003.10.001.
- H. Madsen and J. Holst. Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings*, 22(1):67–79, 1995. ISSN 03787788.
- H. Madsen and J. Schultz. Short time determination of the heat dynamics of buildings. Technical report, DTU, 1993.
- H. Madsen and P. Thyregod. *Introduction to General and Generalized Linear Models*. CRC Press, 2010.
- Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, 2001.
- A. Rabl. Parameter-estimation in buildings - methods for dynamic analysis of measured energy use. *Journal of Solar Energy Engineering-transactions of the Asme*, 110(1):52–66, 1988. ISSN 0199-6231.
- R. C. Sonderegger. *Dynamic models of house heating based on equivalent thermal parameters*. PhD thesis, Princeton Univ., NJ., 1978.